THIS WEEK

EDITORIALS

Risk management

A project to pool data and tools to calculate earthquake hazards is an important milestone, but it will be down to individuals to decide how to interpret and respond to those risks.

The number of people living in earthquake zones is rising year on year, making the mitigation of seismic risk more important than ever. So let's hear two cheers for the consortium of Earth and social scientists and engineers that is set to release a Global Earthquake Model (GEM) in the coming year (see page 290).

Only two cheers? Although the project's worldwide scale matches its ambition, whether it will save lives depends on how its data and tools are used locally. And, as shown by the conviction of six scientists for misreporting the risk of an earthquake that hit L'Aquila, Italy, in 2009, it will also depend on how that information is communicated to decision-makers.

The GEM project addresses a need: around the world most seismic-safety workers must cobble together information from a host of sources to work out which locations are the most dangerous. In countries such as Indonesia and Peru, access to seismic information is limited. The digital platform, global databases and suite of software tools offered by GEM will make it much easier for hazard analysts and emergency planners to assess risk. But that is only the start.

Hazard assessment is a specialist trade and open to misinterpretation. GEM's critics contend that the project's authoritative plots and snazzy graphics might generate a false sense of security in the robustness of the results. Feed in a different set of historical quakes or tweak the parameters, and the maps change. Such difficulties are, rightly, hotly debated in the earthquake-hazard community, and better methods may emerge as a result of bigger studies made possible by GEM. Conveying uncertainties will be essential. But disagreement does not diminish risk.

The sharing of knowledge and best practice could help to tip the balance and persuade governments and communities to take action to improve their building stock, and not view earthquakes simply as 'acts of God' about which little can be done. A worldwide network such as GEM could be a conduit. Its standard tools might bring some sanity to seismic-risk analyses in countries such as Italy, where researchers are cowed, and in Greece, where investment is being diverted away from safety to dubious studies of quake prediction.

To be useful, the data should be as comprehensive as possible. Governments and universities worldwide should embrace transparency and publish and pool their seismic, planning and socio-economic data within GEM.

Translation of the results into action should be a priority. In addition to stacking its boardroom with more managers as its membership grows, GEM should extend its training of scientists and practitioners to spread knowledge of GEM and seismic-risk modelling on the ground, where it matters most. As training workshops under way in several regions already show, a major part of GEM's legacy lies in bringing together previously unpublished local data under one umbrella. Fellowships for students and postdocs to work with GEM, as well as the recruitment and instruction of more trainers, would be good steps.

Once the inevitable happens and a major earthquake strikes, GEM should learn from it. Part of the project's legacy must be for the seismic-hazard community — and those working on other hazards — to evaluate how well its information reaches those on the ground and whether it helps to prevent deaths. Given the large sums of money invested, there needs to be a rigorous assessment process by GEM and independent social scientists after five years or so.

"Governments and universities worldwide should pool their seismic, planning and socio-economic data."

The GEM consortium hopes to stay out of the risk-communications fray. It is possible that some person or government could try to hold GEM accountable for providing information that turns out to be incorrect. But, with the backing of the Organisation for Economic Cooperation and Development, the World Bank and several national governments, GEM is more resistant to litigation than individual researchers or a committee in one country, such as the scientists involved in the L'Aquila case. As ever, it is individuals who will have to make the difficult decisions about how to report and respond to quake risk. GEM's value is in helping practitioners to speak the same language. ■

Brain blast

DIY attempts at electrical brain stimulation to improve cognition are to get easier.

Buyer beware. For US$249 a company in the United States is promising to send curious and competitive players of computer games an unusual headset. The device, the company claims, will convert electronic gamers into electronic-gamers. At the touch of a button, the headset will send a surge of electricity through their prefrontal cortex. It promises to increase brain plasticity and make synapses fire faster, to help gamers repel more space invaders and raid more tombs. And, according to the publicity shots on the website, it comes in a choice of red or black.

The company is accepting orders, but says that it will not ship its first headsets to customers until next month. Some are unwilling to wait. Videos on the Internet already show people who have cobbled together their own version with a 9-volt battery and some electrical wire. If you are not fussy about the colour scheme, other online firms already promise to supply the components and instructions you need to make your own. Or you could rummage around in the garage.

That's 'could' as in 'you might be able to', by the way; not 'could' as in 'it's a good idea'. In fact, to try to boost cognitive performance in this way might be a very bad idea indeed. Would it work? It might or it might not. Nobody knows. All we know for sure is that the technology, known as transcranial direct-current stimulation (tDCS), is likely to soon get into the hands, and onto the heads, of many more people.

Experimentation with electricity to improve human performance is not new. Scribonius Largus, court physician to the Roman emperor Claudius, suggested in AD 46 that a live electric ray could be applied to the head of a patient with a headache. The recent surge in interest in tDCS piggybacks on an increasing number of academic studies of its potential to boost cognitive ability, which themselves build on decades-old work using electrical stimulation of the brain to treat ailments such as depression (see *Nature* **472,** 156–159; 2011).

Nor are unorthodox tests of this technology unusual. When Michael Nitsche, a clinical neurologist at the University of Göttingen in Germany, wanted to investigate a related technique called transcranial magnetic stimulation more than a decade ago, he got permission from university ethics boards but still found a shortage of volunteers. Instead, Nitsche experimented on the brains of himself, his father and his sister.

In an opinion piece published earlier this month, Nicholas Fitz and Peter Reiner of the National Core for Neuroethics at the University of British Columbia in Vancouver, Canada, argue that scientists and regulators can no longer ignore the amateurish meddling with tDCS (N. Fitz and P. Reiner *J. Med. Ethics* http://doi.org/mv8; 2013). "The challenge for the field," they write, "is to develop policy that thoughtfully deals with the issues stemming from people using tDCS devices at home."

Such home use of experimental laboratory kit puts neuroethicists, and journals such as *Nature*, in a bind. To draw attention to it could promote and accelerate its use, and so increase the risk of a mishap. To ignore it leaves the risks unexplored. The scale of at-home tDCS use is unclear at present. It might fizzle out. Or, as scientific interest in the power of electrical stimulation of the brain grows, it might appeal to more enthusiasts, just as the fascination and potential of synthetic biology has spawned a parallel DIY community known as biohackers. The scientific interest is certainly there.

Last month, researchers at the University of Oxford, UK, published a study suggesting that random electrical stimulation of the brain could improve mathematical abilities (A. Snowball *et al. Curr. Biol.* **23,** 987–992; 2013). And there is no lack of exposure. Drawn by the ease of access and the killer copy, science journalists are queuing up to try tDCS for themselves and to write about the effects.

"The scale of at-home tDCS use is unclear at present."

Fitz and Reiner are not the first to raise concerns over the DIY tDCS community. Brain researchers flagged the problem last year, as part of a discussion on the broader ethics of using non-invasive brain-stimulation (R. C. Kadosh *et al. Curr. Biol.* **22,** R108–R111; 2012). The researchers even raised the prospect of the ultimate in pushy parents: those who would use the technology on their children to try to boost their cognitive function. And back in 2011, scientists working on tDCS told *Nature* that they were concerned for the safety of those who tried it at home.

It is easier to raise these questions than to answer them. Fitz and Reiner have some sensible suggestions, ranging from greater reporting of the possible long-term risks of tDCS to mimicking the open communication and education strategy with which the life-sciences field has started to engage biohackers. The first step is to acknowledge the issue to get a sense of how widespread the demand for home electrical self-improvement really is. The next few months will tell us more. ■

Science prevails

The US government gives up its fight to keep age restrictions on the morning-after pill.

A former senior official at the US Food and Drug Administration (FDA), who is older than 50, recently tried to buy the emergency contraceptive Plan B One-Step (levonorgestrel) at 6:30 p.m. on a Saturday evening in a major metropolitan area. She had to go to three shops before she found one with an open pharmacy, necessitating the use of her car. After waiting in line at the pharmacy, she was required to provide proof-of-age identification and her birth date was entered in a computer. Next, the pharmacist walked the medication to the cashier at the front of the drugstore, where she was obliged to wait in a queue again. When she reached the counter, she had to publicly point out that the emergency contraceptive waiting on the shelf behind the cashier was hers.

If it was this challenging, logistically and socially, for a highly educated scientist to obtain the 'morning-after pill', imagine what it is like for a 17-year-old girl, or an undocumented immigrant or a single mother with no car and no driver's licence.

Happily, these obstacles will soon be things of the past. Last week, the administration of President Barack Obama dropped its legal quest to keep in place a requirement that girls younger than 15 years old obtain a doctor's prescription to buy the one-dose pill — which becomes less effective the longer after unprotected intercourse it is taken. The change makes the drug available to anyone, with no proof-of-age requirement, on open shelves, and not behind the pharmacy counter.

If it makes parents queasy to know that 13- and 14-year-olds will now be able to purchase the pill with no questions asked, two things are worth noting. A paper published in April confirms earlier findings that sex in this age group is rare (L. B. Finer and J. M. Philbin *Pediatrics* **131,** 886–891; 2013). The same paper finds that girls aged 14 or younger are less likely than 15-year-olds to use contraception the first time they have sex, and that they take longer than older girls to begin using it. Another study found easy access to the morning after pill does not increase promiscuity in the youngest teenagers (C. Harper *et al. Obstet. Gynecol.* **106,** 483–491 (2005).

Nonetheless, there is little comfort to be drawn from the Obama administration's final climb-down on this issue. In a textbook case of political interference in science, successive foot-dragging administrations have for more than 12 years blocked women's and reproductive-rights advocates' attempts to win over-the-counter status for Plan B.

During that time, FDA staff scientists and expert advisers repeatedly found that the pill met the agency's requirements for over-the-counter status for women and girls of all ages. Yet, in an unprecedented and deeply worrying action, in 2011 the Obama administration, in the person of health and human services secretary Kathleen Sebelius, overruled its own FDA's decision to lift the age restriction.

That same administration walked away from the case last week not because of any change of heart, but because it saw that it was going to lose before a judiciary that, rightly, has called the government's tactics arbitrary and capricious.

The administration's actions and attitude, coming from a White House that has vowed very publicly to back its scientists, and not undermine them, remain disconcerting. They raise concerns for the future independence of the regulatory scientists who are employed to apply science to existing law.

If this administration, or any White House, has a political issue with that law — if, for instance, it wants to enact a bill prohibiting emergency contraceptives for minors — let it do so openly, lobbying for such a measure in Congress. There, and not in the science agencies, is where politics belongs. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

Sharing information is preferable to patenting

The US Supreme Court ruling on gene patents is a welcome boost to efforts to increase the free exchange of scientific information, says **Colin Macilwain**.

The prevailing commercial ethos in the life sciences over the past 30 years has been that academic biologists hold their results close to their chests and keep an eye out for patent opportunities. This approach took root after the 1980 passage of the Bayh–Dole Act, a US law that allowed publicly funded intellectual property to be handed over to universities and private companies without strings, and so gave birth to the biotechnology industry.

Some of the strings were reattached last Thursday, when the US Supreme Court finally said 'no' to human gene patenting — ending a three-decade charade in which the US Patent and Trademark Office (PTO) liberally issued patents on single genes. The court's judgment struck down the patents on breast-cancer susceptibility genes held by Myriad Genetics in Salt Lake City, Utah, but still allows the patenting of synthesized complementary DNA (see page 281). It may make little difference to the patent landscape in the short term. But the 13 June ruling is of great symbolic significance, for it happens to coincide with a general retreat from patenting as the goal and driving force for biological discovery.

I never did understand the PTO's position on such patents. The analogy that worked for me was with the periodic table: patenting a gene from nature is akin to patenting a chemical element from nature, which seems absurd. It turns out that the US Supreme Court agrees.

The decision marks a great victory for patent 'sceptics' such as the Public Patent Foundation, based in New York, which has been fighting the Myriad patents for years. These groups see patent protection as a restriction on the freedom to innovate, rather than a spur to do so.

PATENTING A GENE FROM NATURE IS AKIN TO PATENTING A CHEMICAL ELEMENT FROM NATURE, WHICH SEEMS ABSURD.

The drug and biotechnology industries will continue, of course, to seek patent protection for everything that moves. But the trend I see is one that moves in the opposite direction — towards the free sharing of scientific information and open innovation.

At least three global developments vouch for this. One is the decline of biology's single-laboratory approach, and its growing reliance on large, collaborative groups sharing huge volumes of data. Under this massively collaborative approach — which is closer to how much of engineering and the physical sciences already operates — patenting loses its sway, as everyone relies on everyone else's techniques and ideas.

The second is the fact that the powerful nations in the new world order — India, Germany, Brazil and China — are each, in their own ways, less committed to patent protection than is the United States, whose property-fixated founders even wrote it into the national constitution. The United States twisted these nations' arms to sign the Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement, negotiated in 1994 and implemented in 2001, that committed them to adhere to patent protection along US lines. But as US military and economic dominance fades, so will TRIPS and its consequences.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/oarbzx

The third factor is the rise of the 'open innovation' movement, which is making solid gains. ResearchGate, a Berlin-based information-sharing portal backed by Microsoft founder Bill Gates, has already attracted 2.9 million participants worldwide, most of them working in medicine or biology. And the open-access policy of the London-based Wellcome Trust, the world's largest biomedical research charity, was extended in April to require much Wellcome-funded work to be published with the least restrictive Creative Commons licence, allowing the papers' free dissemination by others — including for commercial gain.

The idea of open innovation is already well entrenched in information technology and other high-tech sectors, where companies find that they can meet customers' needs faster by building on each others' ideas. In many cases, broad cross-licensing agreements sweep patent obstacles out of the way.

The biotechnology and pharmaceutical industries argue that medicines are a special case, and that without patent protection, no-one would bear the costs of obtaining regulatory approval for new drugs and devices. These regulatory barriers are so high in the first place, of course, because of industry's persistent and sometimes reckless attempts to circumvent them.

The patent-based model of innovation in biotechnology, as it stands, does bear occasional fruit in oncology, in which, as a biotech analyst once earnestly informed me, 'successful' drugs will extend a patient's life by six months, at US$10,000 a month. But our most pressing public-health needs are for new antibiotics and treatments for conditions such as Alzheimer's, which will never generate such windfall profits.

In genetic testing, Myriad's model of charging $3,000 for its test will be, it turns out, a bizarre one-off. The future public-health need will be for low-cost, multiple-gene tests, uninhibited by thickets of patents.

Efforts to develop an intellectual-property model that bypasses patents, such as the one proposed by the Biological Innovation for Open Society initiative in 2004 (see *Nature* **431,** 494; 2004), have not progressed very far, and established models of innovation will not be overturned in a day. But they surely will evolve in ways that reflect the interests of the public, which is, after all, paying for the research, the diagnostics and the medicines. ■

Colin Macilwain *writes about science policy from Edinburgh, UK.*
e-mail: cfmworldview@gmail.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

ECOLOGY

Salt water fuels nitrogen release

Saltwater incursions into coastal wetlands can increase the release of ammonium into the ocean, complicating coastal management in the face of human development, climate change and rising sea levels.

Marcelo Ardón at East Carolina University in Greenville, North Carolina, and his colleagues analysed the impact of increased saltwater levels on natural and restored wetlands in North Carolina during four droughts from 2007 to 2012. Reduced processing of ammonium by soil microbes and less nitrogen uptake by plants contributed to ammonium runoff, but releases were higher in restored wetlands, probably due to residual nutrients from fertilizer use.

The researchers suggest that the potential for saltwater-induced nitrogen release should be assessed during the development of large coastal wetland-restoration projects.

Glob. Change Biol. **http://dx.doi.org/10.1111/gcb.12287 (2013)**

TOP LEFT, B. MOOSE PETERSON/ARDEA.COM; BOTTOM LEFT, DE AGOSTINI PL/GETTY

TOP RIGHT, MINDEN PICTURES/SUPERSTOCK; BOTTOM RIGHT, NORBERT WU/MINDEN PICTURES/FLPA

EVOLUTION

Diving is in the blood

Diving mammals ranging from water shrews, beavers and seals to ancient whales (pictured clockwise from top left) share adaptations in the protein that stores oxygen in muscles.

A team led by Michael Berenbrink at the University of Liverpool, UK, analysed the myoglobin proteins of extant mammals, and from this inferred the sequences of these proteins in the mammals' extinct relatives. Compared with non-divers, long-diving creatures tended to have higher levels of myoglobin in their muscles and these proteins were more highly charged, which probably prevents them from sticking together and reducing their utility. On the basis of this relationship, the team developed a model to estimate how long ancient animals could have stayed underwater. They calculated that after ancestors of whales moved from land to water in the Eocene, 56 million to 34 million years ago, their diving capacity increased from 1.6 to 17.4 minutes.

Science **http://dx.doi.org/10.1126/science.1234192 (2013)**

ASTROPHYSICS

Magnetic energy of supernovae

Light from five super-luminous supernovae has revealed an unusual power source behind these cosmic explosions, which were 5 to 100 times brighter than regular supernovae.

Cosimo Inserra at Queen's University Belfast, UK, and his team monitored five nearby supernovae for up to a year each. They report that the persistent glow of these flare-ups spotted by the Panoramic Survey Telescope and Rapid Response System in Maui, Hawaii, is too bright to be generated by radioactive nickel, the fuel of conventional supernovae. However, both the peak luminosity and the long tail of the light are consistent with stars collapsing to form magnetars — rapidly spinning neutron stars with powerful magnetic fields — that provide an additional reservoir of energy for the supernova. This is the strongest observational evidence so far for this supernova mechanism, the researchers say.

Astrophys. J. 770, **128 (2013)**

CLIMATE CHANGE

Acidic waters do not toughen corals

Even corals that have spent generations in acidic waters have failed to adapt completely to these harsh conditions.

As atmospheric levels of carbon dioxide increase, the world's oceans are becoming more acidic, with potentially serious consequences for animals that have carbonate skeletons and shells. Adina Paytan at the University of California, Santa Cruz, and her colleagues collected samples from seven colonies of *Porites astreoides* coral that live in the seas off the Yucatan Peninsula in southeastern Mexico, where groundwater springs have produced low-pH conditions for thousands of years. These corals had lower growth rates and experienced higher predation by boring organisms than seven samples of the coral living just beyond the influence of springs. Despite living in low-pH areas, these reef-forming organisms will not fully adapt to the ocean acidification conditions expected by 2100, say the authors.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1301589110 (2013)**

BIOLOGY

A fluorescent protein from eels

Muscle fibres of the Japanese freshwater eel (*Anguilla japonica*) produce a fluorescent protein, the first to be identified in a vertebrate.

Atsushi Miyawaki and his colleagues at the RIKEN Institute in Wako, Japan, identified the gene that encodes the protein and named it *UnaG*, after unagi, the Japanese word for this eel. When expressed in mammalian cells, the protein produced green fluorescence. UnaG is inactive until it binds to the naturally occurring small molecule bilirubin, a breakdown product of haemoglobin. The team showed that UnaG can be used to measure bilirubin in human serum. It might also be useful as a laboratory tool alongside other widely used fluorescent proteins.

Cell **http://dx.doi.org/10.1016/j.cell.2013.05.038 (2013)**

For a longer story on this research, see go.nature.com/fljtrl

AAAS

ANIMAL BEHAVIOUR

Turtle tots chase warm spots

Cold-blooded turtles move towards the most comfortable climes, even while they are still embryos.

Wei-Guo Du at the Chinese Academy of Sciences in Beijing and his colleagues heated the ends or sides of recently laid eggs of the Chinese pond turtle (*Chinemys reevesii*, **pictured**) for a week and measured the movements of the embryos by shining light through the shells. Embryos moved towards spots maintained at a balmy 29 °C or 30 °C, but shifted away from spots heated to a dangerously hot 33 °C. Only living embryos changed position, suggesting that the motion was due to the animals rather than to changes of viscosity in egg fluids. Although reptile embryos are generally thought to lack control over their environment, turtles inside eggs behave much like adults to regulate body temperatures, the authors say.

Biol. Lett. 9, **20130337 (2013)**

For a longer story on this research see go.nature.com/8ixxah

NATURE PRODUCTION/NATUREPL.COM

MARINE SCIENCE

Marine dumping detailed

Humans are dumping far more litter in the ocean than was once thought.

Kyra Schlining at the Monterey Bay Aquarium Research Institute in Moss Landing, California, and her team used a database of characterized observations from 22 years of research-submersible missions in Monterey Bay to identify anthropogenic marine debris. The litter was seen in 1.49% of the surveyed area, was mainly metal and plastic, ranged in type and depth from a PVC pipe at 25 metres to a plastic bag at 3,971 metres, and was especially prevalent around the submarine Monterey Canyon. Most of the metal and plastic debris was seen below depths of 2,000 metres, suggesting that earlier studies may have underestimated the impact of detritus on deep regions, which are generally poorly observed. Submarine canyons may have trapped and funnelled the debris to depth, the authors suggest.

Deep-Sea Res. I **http://dx.doi.org/10.1016/j.dsr.2013.05.006 (2013)**

COMMUNITY CHOICE

The most viewed papers in science

CHEMISTRY

Catalyst targets spot on carbon ring

HIGHLY READ
on pubs.acs.org in May

A metal catalyst makes sure that molecules join at the correct region of a carbon ring, disrupting the same bond every time.

Chemists have made strides in identifying reactions that disrupt notoriously unreactive carbon–hydrogen bonds, which is a key step when attaching molecules to carbon-ring structures. But it is still challenging to ensure that a molecule joins at the correct region of a ring. Frank Glorius and his colleagues at the University of Münster in Germany have solved this problem for a recalcitrant region on a class of rings known as benzo[*b*]thiophenes. Using palladium on a carbon support, together with copper chloride, the authors developed a catalyst with more than 99% selectivity for the desired hydrogen atom in many cases. The reaction is quite cheap, and withstands air and water.

J. Am. Chem. Soc. 135, **7450–7453 (2013)**

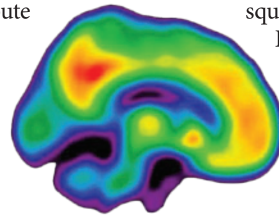
NEUROBIOLOGY

Mutations alter brain amyloid

Some mutations that boost the risk of Alzheimer's may also increase production of a form of amyloid-β, a peptide that is thought to contribute to the disease.

People who inherit specific mutations of the genes *PSEN1* or *PSEN2* nearly always develop a rare form of Alzheimer's disease. Randall Bateman and his team at the Washington University in St. Louis, Missouri, used stable isotope labelling and positron emission tomography to track production of amyloid-β in the brains (**pictured**) of 11 patients who carry PSEN mutations and 12 of their siblings who do not. Those with the mutation produced a long form of the peptide, called amyloid-β42, at a rate 18% higher on average than those without the mutation. Amyloid-β42 is the main component of amyloid plaques, which are found in the brains of patients with Alzheimer's.

Sci. Transl. Med. 5, **189ra77 (2013)**



PALAEONTOLOGY

Early animals' revealing tracks

Fossilized trails left in 560-million-year-old Canadian rocks may be some of the earliest evidence of squirming animals.

Latha Menon at the University of Oxford, UK, and her team studied the disk-shaped impressions left by an organism called *Aspidella* in what was once shallow water in Newfoundland. The authors identified previously overlooked horizontal and vertical rock trails that seemed to be linked with *Aspidella*. They suggest that the marks were made as the animals wormed their way out of sediment, rather than as they passively slid. *Aspidella*, and perhaps other Ediacarans, were probably early animals living underwater, the authors say.

Geology **http://dx.doi.org/10.1130/G34424.1 (2013)**

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

EVENTS

G8 science summit

Five years since their last meeting, science ministers from the G8 countries met last week at the Royal Society in London, for the first time also including the heads of the national science academies. Antimicrobial drug resistance received particular attention as a major global health challenge. In a statement, the group agreed to work on developing quicker diagnostic tests for microbial infections and more targeted treatments. Officials also proposed cooperating internationally to increase access to peer-reviewed, published scientific results.

PIERRE-PHILIPPE MARCOU/AFP/GETTY

Spanish researchers march against cuts

Rallying against the sharp decline in government support for science, Spanish researchers gathered in 19 cities on 14 June. According to the Letter for Science movement, which organized the protests, some 5,000 scientists marched to the economic ministry in Madrid (pictured) to deliver a set of proposals to stop “the ruin of the Spanish science system”. Spain has cut its science budget by 39% since 2009, and eliminated its science ministry in 2011. An estimated one-third of all projects slated for 2013 funding under the National Plan for Research, Development and Innovation — the country’s main science funding scheme — have yet to receive payments. Letter for Science has called for increases in science spending and the creation of an independent science agency. See go.nature.com/htidtx for more.

POLICY

DNA patents

The US Supreme Court has ruled that naturally occurring human genes extracted from the genome cannot be patented. The 13 June decision marks the end of a lawsuit over the validity of gene patents held by Myriad Genetics, a medical diagnostics company based in Salt Lake City, Utah. The genes, *BRCA1* and *BRCA2*, are associated with breast and ovarian cancers. The court noted that ‘synthetic’ DNA, including complementary DNA synthesized from an RNA template, can still be patented. See page 281 and go.nature.com/nlwsud for more.

Japan NIH opposed

Scientists in Japan are resisting the government’s plan to form an agency modelled on the US National Institutes of Health (NIH). Under the proposal outlined in April, Japanese government officials would select research fields and manage budgets for projects aimed at boosting health-related science. Last week, seven major bioscience societies, including the Molecular Biology Society of Japan in Tokyo, circulated an ‘emergency statement’ warning that such a top-down approach would stifle the creativity and motivation of scientists. In a separate statement, a further 54 scientific societies expressed similar concerns.

Plan B for all

A controversial emergency contraceptive will become available without a prescription to women of all ages in the United States. President Barack Obama’s administration said last week that it would drop its legal bid to continue requiring prescriptions for girls younger than 15 who seek to buy Plan B One-Step (levonorgestrel). In April, a federal judge ordered that the ‘morning after pill’ be sold without this restriction, echoing an earlier decision by the Food and Drug Administration (see *Nature* **496**, 138; 2013). Obama’s justice department was appealing the judge’s ruling. See page 272 for more.

Medical malware

Personal medical devices such as pacemakers and cardiac defibrillators should be safeguarded against hacking, says the US Food and Drug Administration. Responding to the increasing use of wireless and Internet-connected medical products, the agency issued draft guidelines on 14 June that would consider cybersecurity in the regulatory-approval process for new devices. Manufacturers should ensure that medical devices are safe from unauthorized access and manipulation, the agency says.

Chimp change

The US Fish and Wildlife Service plans to declare captive chimpanzees (*Pan troglodytes*) in the United States endangered, bringing them under the same designation as their wild counterparts, according to a 12 June proposal. The change could seriously impede the availability of captive chimps for invasive research, because scientists would be required to obtain permits for any invasive studies by showing that the work would contribute to the survival of the species. The proposal, which will be open for public comment for 60 days, comes as the National Institutes of Health in Bethesda, Maryland, considers retiring most of the 360 chimps it owns. See go.nature.com/ppywfs for more.

RESEARCH

Peruvian forests

Annual deforestation in the Peruvian Amazon declined in 2010 and 2011, according to Peru's first comprehensive analysis, released last week. Peru is the second country, after Brazil, to systematically track deforestation with satellite imagery. The country's programme is based on software from the Carnegie Institution for Science in Stanford, California. Annual deforestation averaged 163,000 hectares in 2005–09, up 79% from the average between 2000–05. By 2011, however, deforestation had dropped by nearly 37%.

PEOPLE

KARL SCHOENDORFER/REX FEATURES

Nobel chemist dies

The death earlier this month of Jerome Karle at 94 was reported on 14 June. Karle shared the 1985 Nobel Prize in Chemistry for helping to develop X-ray crystallography, a technique that uses the scattering patterns of X-rays to reveal the three-dimensional structure of molecules. His mathematical methods led to advances in understanding the structure and function of small-molecule drugs and other complex chemical compounds. Karle worked at the US Naval Research Laboratory in Washington DC from 1944 until 2009. He died on 6 June.

FACILITIES

Marine lab hitched

The Marine Biological Laboratory (MBL) in Woods Hole, Massachusetts, has finalized an affiliation with the University of Chicago in Illinois that is slated to begin on 1 July. Neil Shubin, an evolutionary biologist at the university, will facilitate academic collaborations between the institutes. The MBL will remain an independent entity, and the university will provide oversight and resources. See go.nature.com/lukscr for more.

Neutrino factory

A study ordered by the European Commission in 2008 has identified the leading project proposal for high-intensity neutrino research in Europe. In a report presented last week at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, the Neutrino Factory was ranked as the best option among three facilities for testing whether neutrinos and antineutrinos behave differently. Whether the projected multibillion-dollar accelerator will be built remains uncertain. See go.nature.com/pvce6j for more.

FUNDING

Money for Mars

After years of instalments, the European Space Agency (ESA) on 17 June committed to the final payments for the initial stage of ExoMars, a mission to search for signs of life on the red planet. At the Paris Air & Space Show, ESA signed a €643-million (US$857-million) contract with Thales Alenia Space Italy, based in Rome, that covers the cost of an orbiter spacecraft to be launched in 2016 and preparations for a planned follow-up mission in 2018 to place a rover on Mars.

France grant slump

Government auditors in France have expressed concern over the country's falling share of European Union (EU) research grants. In 2012, France garnered 9.5% of EU grants, down from 14.4% in 2007, according to a report released by the auditors last week. The country contributed €6 billion (US$8 billion) to EU research programmes from 2007 to 2012, but won back just €3.42 billion in grants. In May, France approved a strategic plan called France Europe 2020 to better align domestic research priorities with those of EU-wide programmes.

COMING UP

24–29 JUNE
Researchers gather in San Francisco, California, at the 26th International Symposium on Lepton Photon Interactions at High Energies, to discuss advances in particle physics and astrophysics, and to plan for future facilities.
go.nature.com/pdhzqu

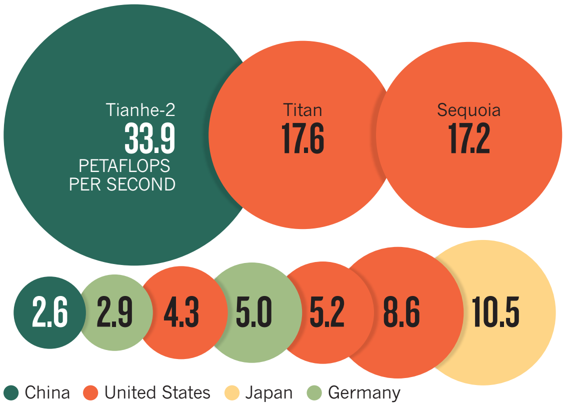
25–28 JUNE
Ghent in Belgium plays host to the International Conference on Genomics in Europe, with next-generation sequencing and bioinformatics on the agenda.
go.nature.com/s6xagp

TREND WATCH

China's Tianhe-2 computer took first place in a list of the world's fastest 500 supercomputers, released on 17 June. With Tianhe-2, to be deployed at the National Supercomputer Centre in Guangzhou, the country reclaims a lead it first gained in November 2010. Over the past 20 years, the power of the leading supercomputer has increased by an order of magnitude roughly every 3.5 years; it should reach an exaflop (10^{18} floating point operations per second, or 1,000 petaflops) by 2019.

CHINA REGAINS FASTEST SUPERCOMPUTER

The top ten most powerful supercomputers were developed in four countries. At 33.9 petaflops per second, Tianhe-2 is the world's fastest.



SOURCE: WWW.TOP500.ORG

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

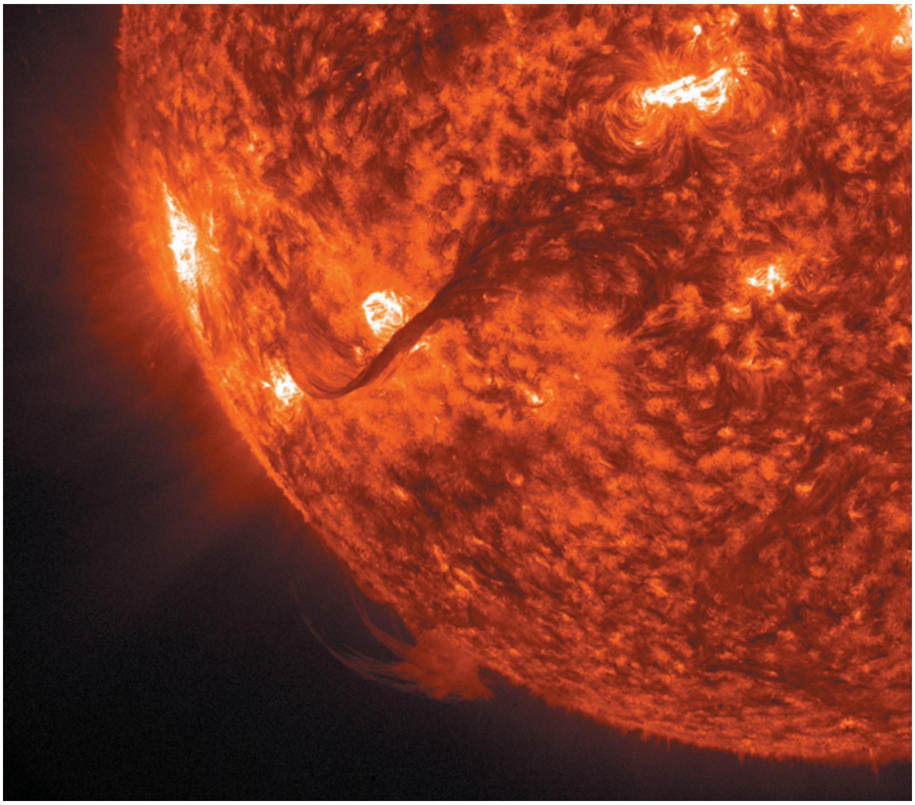
NEWS IN FOCUS

PATENTS Biotechnology companies fret over Myriad court ruling **p.281**

EVOLUTION Research into canine domesticity is a dog-eat-dog enterprise **p.282**

SPACE Opportunity has a way to go to best 1970s Soviet record **p.284**

PHYSICS Credibility grows for quantum-computer company **p.286**



GODDARD SPACE FLIGHT CENTER SCIENTIFIC VISUALIZATION STUDIO/SDO/NASA

The IRIS telescope will zoom in on the chromosphere, pictured here by the Solar Dynamics Observatory.

SOLAR PHYSICS

NASA sets sights on the Sun

IRIS mission aims to scrutinize the layer between the star's surface and its flickering corona.

BY ALEXANDRA WITZE

For the most part, the Sun gets plenty of attention from astronomers. Its surface, or photosphere, bristles with sunspots and erupts with powerful flares. Its outer atmosphere, or corona, shimmers with gossamer arcs mapped out by magnetic field lines. But between these two charismatic regions lies a swathe some 1,700 kilometres thick — the chromosphere — that has largely been overlooked.

This region is about to have its day. On 26 June, NASA plans to launch the US$181-million Interface Region Imaging Spectrograph (IRIS). The instrument's 'eyes', working in the ultraviolet spectrum and designed to follow the flow of matter and energy in the chromosphere, will help astronomers to work out how the photosphere and corona are linked — including how temperatures soar from some 6,000 °C at the solar surface to more than 1 million degrees in the corona. The chromosphere is "a missing piece of the puzzle", says Bart de Pontieu, the IRIS science lead at the Lockheed Martin Solar and Astrophysics Laboratory in Palo Alto, California.

Other missions have looked at the chromosphere before, including the Sunrise 2 high-altitude balloon observatory, which landed on 17 June after a five-day flight in the Arctic. But IRIS will hone in specifically on the chromosphere with greater spatial and temporal resolution than many of its predecessors.

IRIS will take images every five seconds, will obtain spectra every one to two seconds and will be able to discern objects as small as 240 kilometres across. "It's just staggering the dynamics you can see when you have that kind of resolution," says Scott McIntosh, an IRIS co-investigator at the National Center for Atmospheric Research in Boulder, Colorado.

That resolution will help researchers to map out small, finger-like jets of plasma that were discovered in 2007. Using data from the Japanese Hinode satellite and NASA's Solar Dynamics Observatory, de Pontieu and his colleagues tracked these 'type II spicules', which rapidly funnel mass and energy up through the chromosphere and could be a significant factor in the coronal heating problem. But they appeared and died away in minutes — faster than anyone had expected (B. de Pontieu *et al. Science* **331,** 55–58; 2011). "Suddenly it became clear that this new class of events was very important all over the Sun," says Alan Title, a solar physicist at Lockheed Martin and IRIS principal investigator.

IRIS's 20-centimetre telescope and imaging spectrograph are designed to study phenomena that change as rapidly as these spicules. If all goes to plan, an aircraft carrying IRIS will take off from Vandenberg Air Force Base in southern California before releasing a Pegasus rocket that will launch IRIS the rest of the way into space. IRIS will fly in a polar orbit, 660 kilometres above Earth's surface, constantly facing the Sun. Its narrow field of view will be trained on just a small part of the chromosphere: about 1% of the Sun's disk.

➲ NATURE.COM
For more on heliophysics priorities, see:
go.nature.com/vtioeg

First light is ▶

expected about three weeks after launch, with science data arriving several weeks after that. The IRIS team plans to start by answering some long-standing questions about the chromosphere, says McIntosh, such as how many photons are emitted as solar plasma rises up through the chromosphere and how many form as it falls back down, cooling and condensing along the way.

Title says that one of the reasons IRIS is happening now is because modelling work carried out over the past decade has given solar physicists the confidence that they could actually understand the data flowing from a chromospheric mission. The mission team includes modellers such as Mats Carlsson of the University of Oslo, who says that IRIS will help him to understand why his models don't come up with the right amount of heating for the upper chromosphere. "Finally we have some hope of being able to understand these things, by combining simulations and observations," says Carlsson.

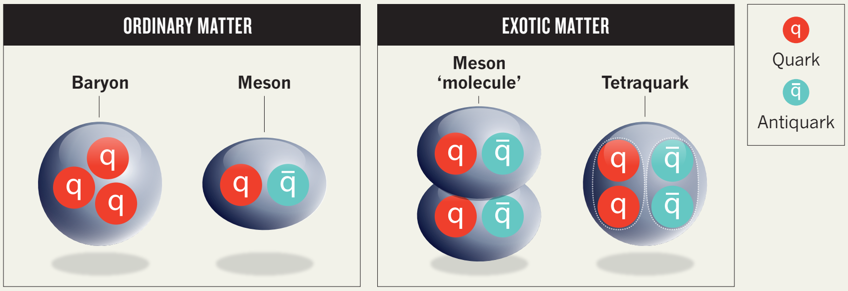
The spacecraft will begin its mission at an opportune time: the Sun is now at the peak of its 11-year cycle of activity, although this peak is much less impressive than the last one. By one measure of solar activity — the amount of radiation emitted by solar storms to reach Earth — the current maximum looks about the same as 1996's solar minimum, says Dean Pesnell, a solar physicist at NASA's Goddard Space Flight Center in Greenbelt, Maryland.

IRIS didn't necessarily have to fly at a solar maximum, but it will be a boon should the Sun flare up during the two-year mission, says de Pontieu. One of the mission's scientific goals is to better understand how kinked magnetic field lines at the Sun's surface trigger big eruptions of matter and energy. IRIS will be able to track these large flares up into the corona, connecting the dots through the earliest phases of a flare's life cycle.

And there is another way in which the mission's timing will be auspicious. In November, the comet ISON is expected to have a close brush with the Sun. IRIS, along with other solar missions, will be in a prime position to watch this happen and could spot unexpected events. There is precedent: in December 2011, a comet named Lovejoy flew through the solar corona, and surprised physicists with the way its waving tail interacted with the Sun's magnetic field. ■

QUARK SOUP

Researchers at colliders in China and Japan have succeeded in making exotic matter comprising four quarks, but are still debating whether the fleeting particles are meson pairs or true tetraquarks.



PARTICLE PHYSICS

Quark quartet opens fresh vista on matter

First particle containing four quarks is confirmed.

BY DEVIN POWELL

Physicists have resurrected a particle that may have existed in the first hot moments after the Big Bang. Arcanely called Z_c(3900), it is the first confirmed particle made of four quarks, the building blocks of much of the Universe's matter.

Until now, observed particles made of quarks have contained only three quarks (such as protons and neutrons) or two quarks (such as the pions and kaons found in cosmic rays). Although no law of physics precludes larger congregations, finding a quartet expands the ways in which quarks can be snapped together to make exotic forms of matter.

"The particle came as a surprise," says Zhiqing Liu, a particle physicist at the Institute of High Energy Physics in Beijing and a member of the Belle collaboration, one of two teams claiming the discovery in papers published this week in *Physical Review Letters*[1,2].

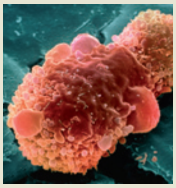
Housed at the High Energy Accelerator Research Organization (KEK) in Tsukuba, Japan, the Belle detector monitors collisions between intense beams of electrons and their antimatter counterparts, positrons. These crashes have one-thousandth the energy of those at the world's most powerful accelerator, the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland, but they are still energetic enough to mimic conditions in the early Universe. Collision rates at KEK are more than twice those at the LHC, and they occasionally give birth to rare particles not found in nature today — ephemeral creatures that wink into existence for an instant and then fall to pieces.

> ***"They have clear evidence of a particle with four quarks."***

Some of that subatomic shrapnel matches what would be expected from the breakdown of a particle containing four quarks bound together: two especially heavy 'charm' quarks and two lighter quarks that give the particle a charge. With 159 of these Z_c(3900) particles in hand, the Belle team reports that the chance that its result is a statistical fluke is less than 1 in 3.5 million[1]. "They have clear evidence of a particle with four quarks," says Riccardo Faccini, a particle physicist

MORE ONLINE

TOP STORY



Study shows how to eliminate false positives in surveys of genes involved in cancer go.nature.com/y7mev5

MORE NEWS

● Self-confessed liar publishes more stem-cell work go.nature.com/otkh5e

● Chemical-weapons experts back case against Syria go.nature.com/zk268h

● Melting under ice shelves is major part of Antarctica's ice loss go.nature.com/uwahaa

NATURE PODCAST

Martian meteorites, D-Wave's quantum computer, and how naked mole rats manage to defeat cancer nature.com/nature/podcast

STEVE GSCHMEISSNER/SPL

at the Sapienza University of Rome.

The new particle has also been vouched for by a second experiment, the Beijing Spectrometer III (BESIII) at the Beijing Electron Positron Collider. BESIII found 307 $Z_c(3900)$ particles, sifted from 10 trillion trillion electron–positron collisions[2].

"This gives credence to all of the other particles that Belle has seen," says Fred Harris, a particle physicist at the University of Hawaii in Manoa and a spokesman for BESIII. In 2008, Belle found another four-quark candidate[3], and in 2011, it saw two other particles that may have been made of four 'bottom' quarks[4] — but no other particle colliders have confirmed those sightings.

No one questions the number of quarks in the latest particle. More controversial is their arrangement, which could have implications for quantum chromodynamics, the theory describing the strong force that connects quarks. Theorists fall primarily into two camps.

One side proposes that the particle is actually a union of two ordinary particles called mesons, which contain one quark and one antiquark. $Z_c(3900)$ particles could be made up of two mesons joined by a loose connection to form a molecule-like structure (see 'Quark soup').

Other theorists have tentatively labelled the new particle a true tetraquark — four quarks stuck together tightly to form a compact ball. Within the ball, two quarks are bound together, as are two antiquarks. Such pairings do not occur in any known particle and would thus introduce new building blocks of matter — with the potential to guide computer simulations aimed at working out all the structures that quarks can form.

Proponents of the tetraquark theory point out that a 'molecule' made of mesons should split easily into two halves, and that such a breakdown has not appeared in the data. "The signature of a molecule is not seen, which favours the tetraquark picture," says Ahmed Ali, a particle physicist at DESY, Germany's high-energy physics laboratory in Hamburg. But the experiments' margin of error is still too great to rule out the possibility of molecular mesons breaking down. Another way to test the two theories would be to look for other particles that each predicts should exist.

Hoping to end the debate, researchers at BESIII are continuing to dig through data collected since their first experimental run in December and January. Depending on what they find, the unmasking of $Z_c(3900)$ may have to wait for the new, more powerful version of the Belle detector planned to come online in 2015. ■

1. Liu, Z. Q. *et al. Phys. Rev. Lett.* **110,** 252002 (2013).
2. Ablikim, M. *et al. Phys. Rev. Lett.* **110,** 252001 (2013).
3. Chen, K.-F. *et al. Phys. Rev. Lett.* **100,** 112001 (2008).
4. Adachi, I. *et al.* Preprint at http://arxiv.org/abs/1105.4583 (2011).

BIOTECHNOLOGY

Myriad ruling causes confusion

Change to gene patents leaves US biotech in a lather.

BY HEIDI LEDFORD

There was a party waiting for Elizabeth Chao when she arrived for work last week at Ambry Genetics, a medical-diagnostics company in Aliso Viejo, California. On 13 June, the US Supreme Court ended the 30-year-old practice of awarding patents on human genes — an outcome that Chao, a geneticist and chief medical officer of Ambry, had wanted for a long time. "It's such a win for patients," says Chao. "Everyone was crying, jumping up and down and shouting."

In Washington DC at law firm Sughrue Mion, patent lawyer William Simmons was having a rather different day, fielding phone calls from agitated clients in the US biotechnology industry. Although the Supreme Court case was limited to human DNA, the ruling will probably be applied to other molecules such as proteins, as well as to other organisms — including agriculturally important plants. "It's a mess," says Simmons. "We had a lot of clients saying, 'What are we going to do?'"

The Supreme Court decision ended a long-running, emotionally charged legal challenge to gene patents held by Myriad Genetics, a genetic-testing company in Salt Lake City, Utah, on two cancer-associated genes: *BRCA1* and *BRCA2*. The court's first point rang clear — that naturally occurring human genes cannot be patented — and seems poised to broaden the genetic-testing market (see 'Competitors stake their claims').

Yet the grey area between this ruling and the court's second point — that patents can be claimed on modified DNA — has puzzled observers. Gene patent holders, including Myriad, had long argued that the mere act of isolating a piece of DNA from a genome was enough modification for a patent claim, because isolation requires severing the chemical bonds that tether the gene to its surroundings. The Supreme Court justices — and many scientists — disagreed. But patent lawyers are now tearing their hair out over the issue of how much modification is enough. "They've created this bizarre rheostat about the amount of change that would need to take place chemically in order to justify a patent," says Simmons.

Some of the confusion stems from how the Supreme Court justices defined the term synthetic DNA. The court seemed to use it to refer to DNA that had been modified ▶

GENE TESTING

Competitors stake their claims

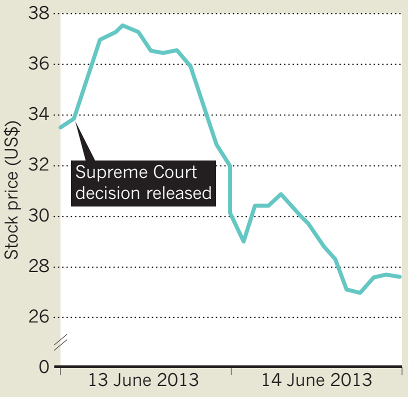
Despite the US Supreme Court decision, Myriad Genetics still has more than 500 claims covering its tests for breast-cancer genes. This reassured investors, and its stock price soared (see 'Patently unclear').

But the price later slumped as firms including California's Ambry Genetics and Quest Diagnostics in Madison, New Jersey, said that they plan to launch their own tests. Myriad's test costs US$4,000, but competitors' tests could cost half that.

Some lawyers say that firms are risking lawsuits by infringing Myriad's patents, but others think that the competitors could design genetic tests that avoid the remaining narrow claims. Any lawsuits could last for years — by which time many of Myriad's patents will have expired. **H.L.**

PATENTLY UNCLEAR

The stock price for Myriad Genetics initially rose after the US Supreme Court ruled against five of its human gene patents, but has since fallen.



SOURCE: NASDAQ

by the inventor. It also gave explicit patent protection to a modified form of DNA called complementary DNA (cDNA), which is made in the lab with an enzyme that creates DNA using an RNA template. Patents on cDNA are deemed more commercially valuable than patents on naturally occurring genes, in part because cDNA tends to be shorter and easier to work with in the lab than genes in their natural state. It can also be used for diagnostic tests if the mutations of interest are contained within the RNA template, as is often the case. But patents on cDNAs, at least for known genes, are largely a dying breed because making cDNA is a common practice that would be considered too obvious for a robust patent.

Increasingly, scientists define synthetic DNA as that which has been made from scratch by assembling the individual bases of DNA into a given sequence, often using machines. And the justices did not say whether synthetic DNA of this sort could be patentable if it exactly copied a naturally occurring sequence.

Lawyer Patrick Waller, of Boston firm Wolf Greenfield in Massachusetts, says that the decision could jeopardize patents on short stretches of synthetic DNA that are used to check whether the genome contains certain sequences, or to create multiple copies of particular DNA regions.

These issues now fall to the lower courts and to patent examiners who must interpret the Supreme Court opinion. Shortly after the decision was issued, Andrew Hirshfeld, a deputy commissioner at the US Patent and Trademark Office in Alexandria, Virginia, issued a memo suggesting that such patents would no longer be granted. That memo, intended to serve as interim guidance until the office updates its policies to incorporate the new ruling, is a sign that the patent office will be interpreting the Myriad decision strictly, says David Berry, a professor of intellectual-property law at the Thomas M. Cooley Law School in Lansing, Michigan. "Companies are just going to have to think up different approaches to claiming their inventions," he says.

Biotech companies might already be changing their approach. Simmons now tells clients to protect certain inventions as trade secrets, which are not publicly disclosed, rather than as patents. After the Myriad decision, he says, he may also instruct clients to introduce many modifications to the DNA or proteins they intend to patent, to make them as different as possible from naturally occurring forms. "There's no guidance here as to what is a sufficient amount of change to warrant a patent," says Simmons. "It's insane." ■

SEE WORLD VIEW P.273

➲ NATURE.COM
For more on gene patents, see:
go.nature.com/kvvjdj

LES HIRONDELLES/FLICKR/GETTY

The issue of when or where canines were domesticated has geneticists in a tug of war.

EVOLUTION

Dog genetics spur scientific spat

Researchers disagree over canine domestication.

BY EWEN CALLAWAY

Scientists investigating the transformation of wolves into dogs are behaving a bit like the animals they study, as disputes roil among those using genetics to understand dog domestication.

In recent months, three international teams have published papers comparing the genomes of dogs and wolves. On some matters — such as the types of genetic changes that make the two differ — the researchers are more or less in agreement. Yet the teams have all arrived at wildly different conclusions about the timing, location and basis for the reinvention of ferocious wolves as placid pooches. "It's a sexy field," says Greger Larson, an archaeogeneticist at the University of Durham, UK. He has won a £950,000 (US$1.5-million) grant to study dog domestication starting in October. "You've got a lot of big personalities, a lot of money, and people who want to get their *Nature* paper first."

In January, Erik Axelsson and Kerstin Lindblad-Toh, geneticists at Uppsala University in Sweden, and their colleagues reported in *Nature*[1] that genes involved in the breaking down of starch seemed to set domestic dogs apart from wild wolves. In the paper and in media interviews, the researchers argued that dog domestication was catalysed by the dawn of agriculture around 10,000 years ago in the Middle East, as wolves began to loiter around human settlements and rubbish heaps (see *Nature* http://doi.org/mv4; 2013).

But Larson, who has worked with Lindblad-Toh on other projects, says that their claim is dubious. He notes that bones that look similar to those of domestic dogs predate the Neolithic revolution by at least several thousand years, so domestication must have occurred before then. "Why waste space [in a paper] saying something that is patently untrue?" he says.

Axelsson concedes that the changes in starch digestion in dogs could have occurred after they were domesticated. But he also counters that the Neolithic era lasted for thousands of years, and that dogs may have been domesticated during the earliest steps towards agrarian life — when human hunter-gatherers settled down and began eating more starch-rich wild plants.

A second study, published last month in *Nature Communications*[2], argues that dogs were domesticated 32,000 years ago when they began scavenging with Palaeolithic humans in southern China. A team led by Ya-ping Zhang at the Kunming Institute of Zoology in China drew that conclusion from studying the whole genomes of several grey wolves, modern European dog breeds and indigenous Chinese dogs.

But Larson says that there is no evidence to suggest that wolves ever lived in southern China, "so how do you domesticate a wolf if there aren't any?" And Jean-Denis Vigne, an archaeozoologist at the National Museum of Natural History in Paris, agrees, noting that in earlier work, Zhang's team "completely ignored what has been published, even in the frame of genetics".

Peter Savolainen, a geneticist at the KTH Royal Institute of Technology in Solna, Sweden, who co-authored the *Nature Communications* paper, argues that Chinese scientific literature suggests that wolves did once live south of China's Yangtze River, but have since become extinct. But he acknowledges that the date that his team reported — like all molecular dating efforts — relies on several assumptions, such as the number of genetic mutations that develop in each generation.

A third paper[3] argues that a more probable date for domestication was 11,000–16,000 years ago. Posted to the arXiv preprint server on 31 May, the study, like Zhang's, compares the whole genomes of wolves and dogs. But the paper paints an even murkier picture, suggesting that wolves and the ancestors of modern dogs continued to breed together long after domestication, and that the wolf population that gave rise to dogs is extinct.

The authors, a team of geneticists co-led by John Novembre at the University of Chicago in Illinois, declined to comment on their work because it has not yet been published in a journal. But Larson and others say that the paper makes a strong point — that studying the genomes of long-dead dogs and wolves is the only way to settle the dispute. At least three other teams in the United States and several others in Europe are racing to sequence ancient dog and wolf genomes, but researchers say that many specimens will be needed to build a clearer picture of domestication. Still, "we're not in a position to be picky", says Adam Boyko, a dog geneticist at Cornell University in Ithaca, New York, who was involved in the arXiv paper[3]. "We're sort of going to be limited to which samples we can get DNA out of."

The move to look at ancient DNA could make the small field of dog genetics even pricklier, because archaeological bone samples are so precious. Novembre says that he finds the field more fractious than human genetics, and says that his experience has given him pause about future canine work. "It's really intense in the dog world," he says. But Boyko, who also collaborates with the Chinese group, says that although the field is competitive, it remains collegial. "At the end of the day, we sit back and enjoy a beer together when we see each other." ■

[1] Axelsson, E. *et al. Nature* **495,** 360–364 (2013).
[2] Wang, G.-D. *et al. Nature Commun.* http://dx.doi.org/10.1038/ncomms2814 (2013).
[3] Freedman, A. H. *et al.* Preprint available at http://arxiv.org/abs/1305.7390 (2013).

PHARMACEUTICALS

China drugs head fired over article row

Researcher stands by results despite demand for retraction.

BY DAVID CYRANOSKI

Jingwu Zang says he is baffled by the whole affair. Until last month, he was head of a neurodegenerative-disease research unit in Shanghai, China, for London-based drug firm GlaxoSmithKline (GSK). On 22 May, as he tells it, his boss told him that there would be an investigation. The next day, Chinese lawyers showed up at the company to interview him. On 31 May, he was told to hand in his computer and company credit card, and was escorted to his car. "Within a few minutes, I was outside the facility I built," he says.

On 9 June, he received a letter informing him of his official termination of employment.

The investigation has focused on a paper published in *Nature Medicine* that Zang co-authored on multiple sclerosis (MS), his speciality (X. Liu *et al. Nature Med.* **16,** 191–197; 2010). GSK is asking for the paper to be retracted; Zang stands by the results. The Chinese blogosphere is abuzz over the dispute, wondering what it signals for a centre seen as a bellwether for China's budding drugs industry.

Zang set up the global research and development centre in Shanghai in 2007. The centre was considered bold: of the many international pharmaceutical giants that had opened research operations in China in the previous five years, only GSK had given its branch wide autonomy, with control over global operations for an entire development sector, that of neurodegenerative diseases. "In Shanghai, we can make decisions that drive global studies," says Zang.

Now with some 400 scientific staff, the centre has several candidate neurodegenerative drugs in phase I and II clinical trials, Zang says, and he was eager to get one through phase III, to "demonstrate that we can do great science and move a clinical compound forward".

Four years ago, Zang's group started work on a protein called the interleukin-7 receptor (IL-7R). "It was a really exciting story," he says. IL-7R sits on the surface of certain immune cells, and a genetic variant of it had been linked to MS. Nobody knew what the underlying mechanism was, but Zang had a hypothesis — that the IL-7 pathway played a part in the pathogenic expansion of T-helper 17 (T_H17) cells, immune cells that, when present in excess, are thought to contribute to MS.

In 2010, the group published results in *Nature Medicine* concluding that this was indeed the case. But last month, the paper came under scrutiny from within GSK after the company and *Nature Medicine* were notified of a problem with some of the data. A GSK investigation has since concluded that human blood samples used to create a figure in the article — described in the caption as being taken from patients with MS — actually came from healthy subjects.

On 10 June, GSK posted a statement saying: "Regretfully, our investigation has established that certain data in the paper were indeed misrepresented. We've shared our conclusion that the paper should be retracted and are in the process of asking all of the authors to sign a statement to that effect."

Zang and Xuebin Liu, the paper's first author, both say that this was an unintentional mistake that does not change the paper's overall conclusion. Liu, whose group ran the experiment and compiled the data, says that the team had hoped to use data from cells of patients with MS and had drafted a manuscript with that wording. But although preliminary data from patients did reveal Zang's proposed link between the IL-7R pathway and T_H17 cells, staining in those images was inadequate — so the team turned to healthy subject data instead, Liu says. In a hurry to beat competition, they forgot to change the caption. Liu says that cells from either group can be used to show the effect.

> **"Regretfully, our investigation has established that certain data in the paper were indeed misrepresented."**

Liu also addressed another problem, noted later on a pharmaceutical blog, after news of the investigation came out: two images, with captions describing different experimental conditions, are identical. Liu says that the mistaken duplication occurred during editing and layout of the article, and has asked *Nature Medicine* to check. The journal's chief editor Juan Carlos López says that he cannot comment yet.

The main thrust of the paper — that IL-7R is related to MS, and that blocking its function can ameliorate MS-like disease in a mouse model — largely agrees with results from other groups. But scientists have failed to replicate the specific mechanism proposed by Zang's team.

One of those studies, led by researchers at Stanford University in California and at Rinat, a subsidiary of the drugs giant Pfizer based in South San Francisco, California, found

▶

that the effects of blocking IL-7R largely agreed with the results from Zang's group. But they were unable to reproduce the results of Zang and Liu's experiments that supported a connection between IL-7R and $\mathrm{T_H}17$ cells as the mechanism. The discrepancy was "not likely due to differences in the experimental protocols, because we diligently followed their methods", the researchers wrote in their 2011 report in *Science Translational Medicine* (L.-F. Lee *et al. Sci. Transl. Med.* **3,** 93ra68; 2011). Stanford's Lawrence Steinman, a corresponding author on the California study, declined to comment.

But Liu says that the California experiment differed from the China experiment in an important way: the GSK team used mature $\mathrm{T_H}17$ cells whereas the California group used undifferentiated ones. "It's a different protocol, a different stage," he says.

Liu says that as first author he takes full responsibility for the mistakes. On 9 June, he announced his resignation on a Chinese bioscience website. Both Liu and Zang say that they stand by the paper's results and will not sign a letter to *Nature Medicine* requesting that the paper be retracted. López says that a retraction is still possible, even if not all of the authors agree to it, "if confidence in the paper is lost". In such cases, "the paper is still retracted, explaining who agrees and who doesn't agree to the retraction", he says.

Asked whether a retraction is warranted if the mistakes do not affect the paper's findings, David Daley, GSK's director of global external communications, acknowledges in an e-mail that in the period since the research was carried out, "an independent body of evidence has accumulated that the receptor [for] interleukin-7 (IL-7) is a valid target for a variety of autoimmune disorders". But he adds that "because certain data in the publication were misrepresented, we believe retracting the paper is the only appropriate action to take".

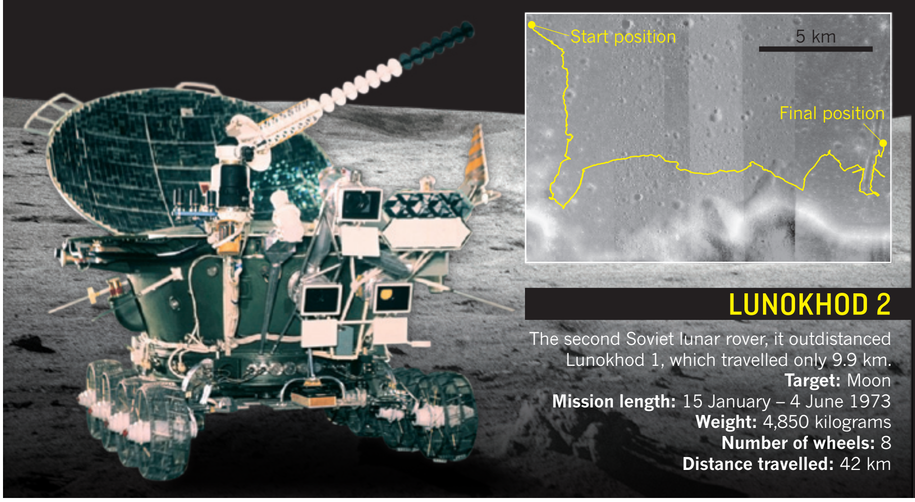
Zang, who was involved in the experimental design and in drafting the manuscript, but not in the hands-on experiments or data compilation, believes that he was fired not over data misrepresentation but for allegedly "influencing the investigation". The letter terminating his employment states that he "wilfully and purposefully undermined and misled this investigation and provided untruthful information" — charges he vehemently denies.

Daley declined to detail GSK's reasons for terminating Zang's employment, but provided a GSK statement: "We are confident in the thorough investigation we conducted and the actions we have taken as a result of our findings. We will not tolerate activity and behaviour that falls short of the high standards expected from our employees."

Zang says the whole episode is bizarre. "I still can't understand it." ■

SPACE RACE

The Soviet moonwalker Lunokhod 2 travelled 42 kilometres, 5 km farther than scientists had long thought — delaying a chance for NASA's Mars rover Opportunity to set a new off-Earth driving record.



LUNOKHOD 2

The second Soviet lunar rover, it outdistanced Lunokhod 1, which travelled only 9.9 km.
Target: Moon
Mission length: 15 January – 4 June 1973
Weight: 4,850 kilograms
Number of wheels: 8
Distance travelled: 42 km

EXPLORATION

Space rovers in record race

Revised data show Soviet Union's 1970s lunar vehicle outdistanced NASA's Opportunity — for now.

BY ALEXANDRA WITZE

Alexander Basilevsky always wanted to stop driving. As a planetary geologist working with the Soviet Union's remotely controlled lunar rovers — Lunokhod 1 and Lunokhod 2 — in the early 1970s, Basilevsky was constantly asking mission chiefs to halt the rolling explorers for scientific studies, fascinated by the buffet of rocks and soil captured by the vehicles' cameras. But the bosses in the Soviet space programme were having none of it. "It is Lunokhod, not 'Lunostop'!" they told Basilevsky as they kept the rovers driving, intent on covering as much ground as possible.

Now it seems that the second rover, Lunokhod 2, went even farther than many back then had thought. New calculations, using images from orbit that trace the rover's 40-year-old tracks far below, show that Lunokhod 2 travelled some 42 kilometres in its lifetime — 5 kilometres more than the distance recorded in the official mission logs. And that means that NASA's Opportunity rover, inching up to the 37-kilometre mark after nearly a decade on Mars, has a long way to go to break the record for the distance driven by a wheeled vehicle on another world (see 'Space race').

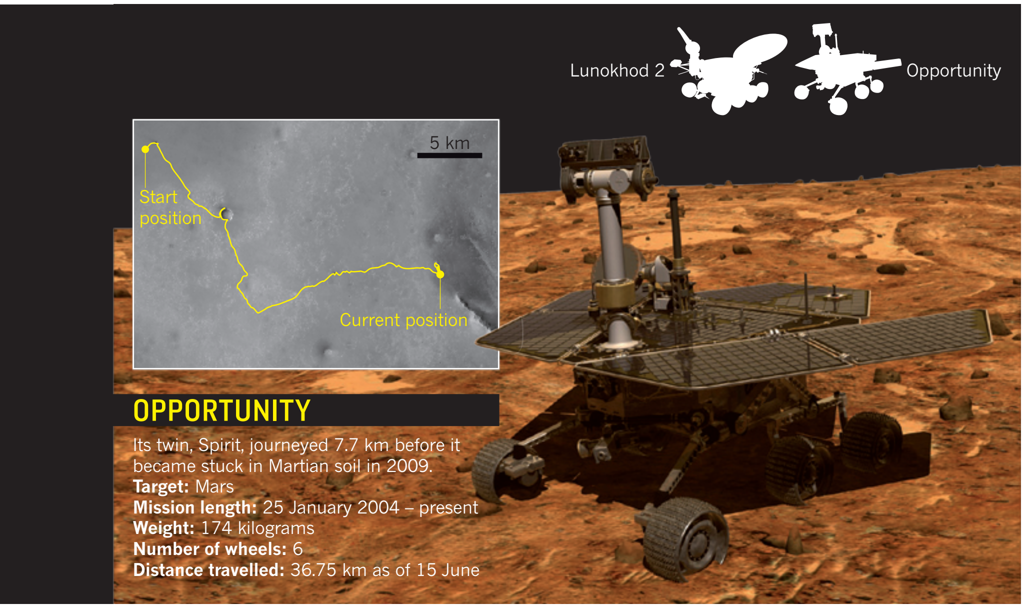
In a mid-May news release about Opportunity's longevity, NASA cited the 37-kilometre distance for Lunokhod 2, and some team members speculated to the press that Opportunity would soon set a new record for driving distance off-Earth. Since then, they have pulled back from any predictions of besting the Russians, even though Opportunity's odometer was at a tantalizing 36.75 kilometres as of 15 June.

"We're not going to talk about breaking any records" just yet, says Opportunity's principal investigator, planetary scientist Steven Squyres of Cornell University in Ithaca, New York. "I'm awestruck by what the Lunokhod team managed to accomplish so many years ago, and I wouldn't want to claim that we've passed their record unless we're really sure."

Russian scientists, for their part, are quite certain about their revised 42-kilometre estimate, and have reported the findings at various planetary-science conferences over the past year.

The 1.7-metre-long Lunokhod 2 rover explored the Moon's

➲ NATURE.COM
For more on NASA's Mars rovers, see:
go.nature.com/fod9ud



OPPORTUNITY

Its twin, Spirit, journeyed 7.7 km before it became stuck in Martian soil in 2009.
Target: Mars
Mission length: 25 January 2004 – present
Weight: 174 kilograms
Number of wheels: 6
Distance travelled: 36.75 km as of 15 June

LUNOKHOD 2: EDGAR D. MITCHELL/GRIN/NASA; RIA NOVOSTI/SPL; MIIGAIK. OPPORTUNITY: JPL-CALTECH/NASA; NASA/JPL-CALTECH/MSSS/NMMNH

Le Monnier Crater for about 4 months, sending back 86 panoramic pictures and more than 80,000 television images. It stopped working in the spring of 1973, possibly after a close shave involving a crater wall dumped lunar soil into its interior.

The revised calculations of its journey were made by planetary mapper Irina Karachevtseva and her colleagues at the Moscow State University of Geodesy and Cartography (MIIGAiK). The team used images of the Lunokhod 2 landing site collected by the Lunar Reconnaissance Orbiter (LRO), which has been studying the Moon since 2009. They adjusted tiny line-of-sight distortions in those images using a three-dimensional representation of the Moon's topography that was made from LRO laser mapping. Tracking the rover's traverse on these adjusted images yielded the current best estimate of between 42.1 and 42.2 kilometres — very close to the distance of a marathon, the team notes.

Karachevtseva says that she is not surprised that the official mission logs are some 5 kilometres off the latest estimate. Lunokhod 2's odometer was a narrow ninth wheel that dragged behind it as it travelled, notching up distance by how much the wheel spun round. It was always thought to have had an error of 10–15%, she says — in fact, one member of the Lunokhod team who helped to drive the rover told MIIGAiK scientists that the team always thought the distances were underestimated.

The MIIGAiK team also reanalysed the path of the first rover, Lunokhod 1, which explored the Moon in 1970–71. Here, surprisingly, Karachevtseva says the team found that Lunokhod 1 had stopped short of the distance shown in the official mission logs: it covered 9.93 kilometres rather than the recorded 10.54 kilometres. A paper on the Lunokhod 1 findings is in the press at *Planetary and Space Science*, and the MIIGAiK team is finalizing a publication on Lunokhod 2.

It is unclear why the Lunokhod 1 distance was originally overestimated and that of Lunokhod 2 underestimated, says Phil Stooke, a planetary cartographer at the University of Western Ontario in London, Canada. He speculates that Lunokhod 1 might have failed to account for wheel slip, a common problem on powdery lunar soils, whereas Lunokhod 2 might have overcompensated or had some other sort of sensor error.

Wheel slip continues to bedevil rovers on other worlds. Opportunity's twin on Mars, the Spirit rover, slipped more than expected as it climbed Husband Hill, in the Gusev Crater region of Mars. However, as it went downhill, the wheels gained traction such that the total slip was close to zero when its journey was completed (R. Li *et al. J. Geophys. Res.* **113,** E12S35; 2008).

Engineers working on Opportunity calibrate the distance it has covered by reconciling its wheel odometry daily with orbital images, says Ron Li, a Mars-rover mapper at Ohio State University in Columbus. Opportunity is currently leaving an area called Cape York, which it explored for 20 months, and heading towards Solander Point about 1.3 kilometres away, where it will try to keep working through the upcoming Martian winter. Lunokhod 2 may thus hold the extraterrestrial driving record for quite a bit longer.

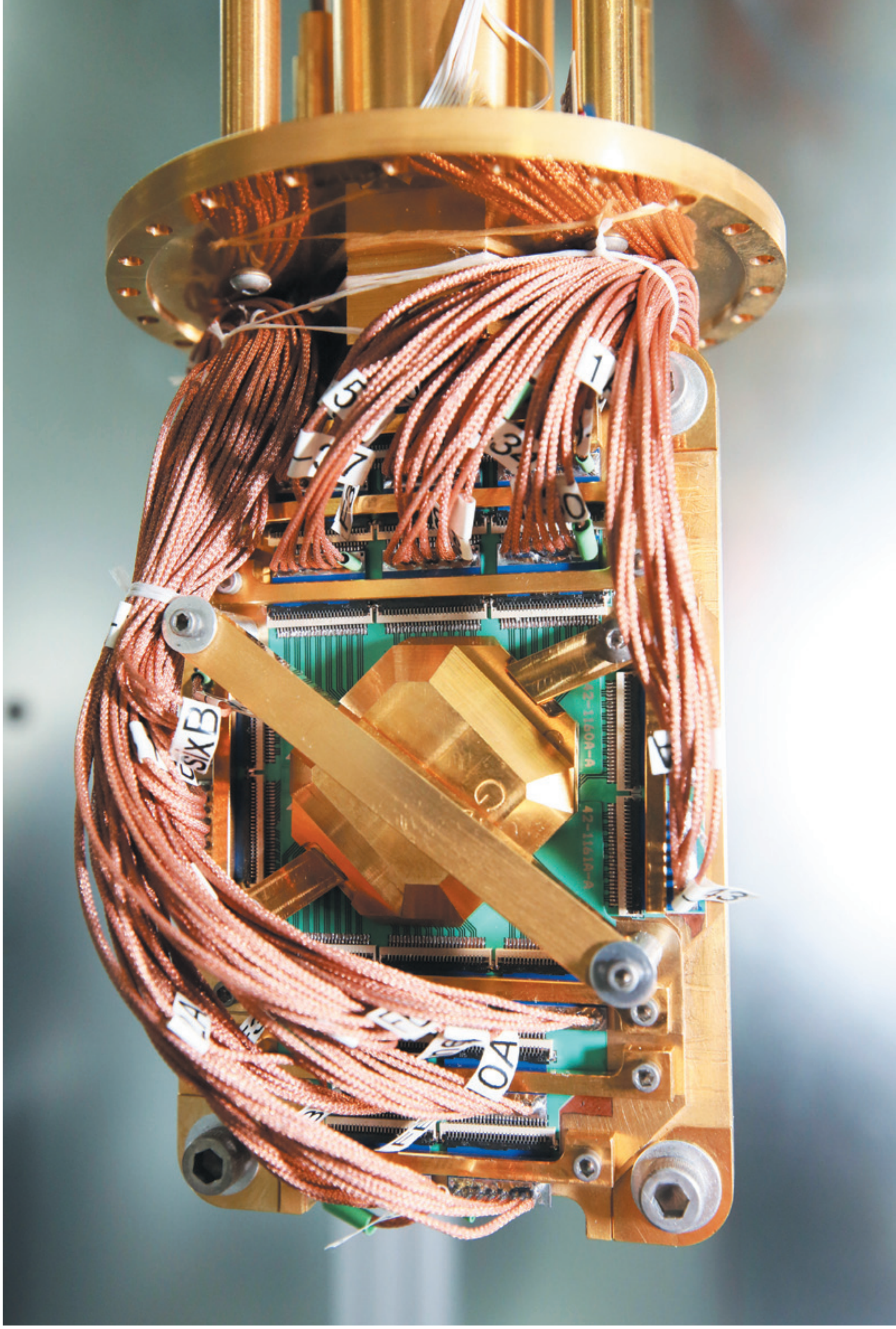
For Basilevsky, now at the Russian Academy of Sciences in Moscow, the reanalysis is a fitting end to the Lunokhod story. As a scientist, he was not supposed to be in the military's mission control centre for Lunokhod 2. But he sneaked in to be present as the rover drivers navigated the alien terrain — and he likes to joke that he drove the machine remotely as well.

"This news," he teases, "says that I was a more effective driver than I used to think." ■

THE QUANTUM COMPANY

D-Wave is pioneering a novel way of making quantum computers — but it is also courting controversy.

BY NICOLA JONES



The D-Wave quantum computer processor is 3,600 times faster than classical computers at some tasks.

KIM STALLKNECHT/THE NEW YORK TIMES/REDUX/EYEVINE

"I've been doing combative stuff since I was born," says Geordie Rose, leaning back in a chair in his small, windowless office in Burnaby, Canada, as he describes how he has spent most of his life making things difficult for himself. Until his early 20s, that meant an obsession with wrestling — the sport that, he claims, provides the least reward for the most work. More recently, says Rose, now 41, "that's been D-Wave in a nutshell: an unbearable amount of pain and very little recognition".

The problem of lack of recognition is fast disappearing for D-Wave, the world's first and so far only company making quantum computers. After initial disbelief and ridicule from the research community, Rose and his firm are now being taken more seriously — not least by aerospace giant Lockheed Martin, which bought one of D-Wave's computers in 2011 for about US$10 million, and Internet behemoth Google, which acquired one in May.

But the pain has been real — much of it, critics would argue, brought on by Rose himself. In 2007, his company announced its first working computer with a showy public demonstration at the Computer History Museum in Mountain View, California. By the current standards of quantum computing — which in theory offers huge advances in computing power — the device's performance was astonishing. Here was a prototype searching a database for molecules similar to a given drug and solving a sudoku puzzle, while the best machines built using standard quantum approaches could at most break down the number 21 into its factors[1].

Sceptics bristled at the 'science by press conference' tone of the introduction, and wondered whether the D-Wave device wasn't just a classical computer disguised as a quantum one. "This company from Canada popped out of nowhere and announced it had quantum chips," says Colin Williams, who published one of the first texts on quantum computing in 1999, and who joined D-Wave last year as business-development director. "The academic world thought they must be crazy."

Today, those criticisms have been quietened to some degree by the release of more details about D-Wave's technology. But they have been replaced by subtler questions: even if the D-Wave computer is harnessing quantum powers, is it really faster or better than a conventional computer? Will it ultimately crack problems that currently take computers decades or more to solve? Or will its capabilities hit a wall?

UNIVERSAL VISION

When Rose founded D-Wave in 1999, he had an engineering degree, a few years' progress towards a PhD in theoretical physics at the University of British Columbia in Vancouver — and no idea how to build a quantum computer. He did have inspiration, from a class on entrepreneurship that he had taken with Haig Farris, one of Canada's best-known technology venture capitalists. Business, says Rose, "appealed to me as being harder than physics or math. There's no prescription for making people do what you want."

Williams' then-new textbook helped to convince Rose that quantum computing would make a suitable target for a new venture. A cheque for Can$4,059.50 (US$3,991) from Farris let him buy a laptop and printer to produce a business proposal. By the early 2000s, D-Wave had attracted millions of dollars in capital, which Rose invested in 15 different research groups to look for the best technology to pursue. "I was like an evangelist, pitching the vision" of a quantum computer, he says.

At the heart of that vision was quantum computing's promise to solve otherwise-intractable problems by drastically reducing the time required to find an answer. The quintessential example is factorizing: like splitting 21 into 3×7, but with numbers hundreds of digits long. That is the basis of the encryption algorithms widely used to protect digital data. Encryption security rests on the fact that conventional computers have to look at every possible factor in turn — a process that takes exponentially longer as the numbers get bigger.

The bottleneck arises because conventional computers store and process information in an either–or fashion, using 'bits' that can each exist in only one of two states, denoted 1 or 0. In most modern computer chips, each bit is represented by the presence or absence of an electric charge. Quantum computers, by contrast, exploit the fuzzy world of quantum mechanics by using 'qubits' that can exist as both 1 and 0 at the same time. In principle, they can explore different solutions simultaneously — reducing a multi-year calculation to seconds.

By the time Rose began his search for the right technology with which to build a quantum computer, researchers had begun to make qubits from many physical systems, including photons that encode zeroes and ones in the direction of their polarization, and ions that encode them in their electron states. They were also working on ways to combine and manipulate the quantum information carried by these qubits, in much the same way that transistor logic gates manipulate the flow of bits in a conventional computer. The goal was to produce 'universal' quantum computers that could carry out any conceivable computation, like a modern classical machine.

➲ **NATURE.COM**
To read another take on quantum computing, see:
go.nature.com/hbsrsz

But this model entailed some huge engineering challenges — starting with the fact that quantum bits are extremely susceptible to outside interference. They are like pencils balanced precariously on their points: the slightest perturbation can knock them off balance, causing an error in the calculation. If each qubit is 99% accurate, an operation involving 10 of them will yield the right answer only 90% of the time, and one with 100 qubits will do so only about 36% of the time. Yet practical applications might require thousands or millions of qubits.

To compensate, developers go to great lengths to shield their qubits from noise, and to devise clever error-correction schemes. But then and now, says Andrew Landahl, who works on quantum computing at Sandia National Laboratories in Albuquerque, New Mexico, "if you look at the redundancy and fidelity you need, it's extremely demanding". Like a rocket that requires tonnes of fuel to hoist a tiny payload, a gate-model quantum computer might need billions of error-correcting qubits just to get 1,000 functional qubits to do something productive.

> "I think it is not too strong to say they were initially ridiculed by the academic community."

By 2003, Rose was convinced that this model was "just a bad, bad, bad idea", he says. So he shifted his focus to what was then a research backwater: adiabatic quantum computing[2]. This technique is best suited to optimization problems — the kind in which the best possible outcome must be found for a number of criteria simultaneously. Examples include trying to arrange the seating for a wedding at which some guests are best friends and others sworn enemies; or finding the most energetically stable way to fold a protein in which the various amino acids attract or repel each other.

All the possible solutions to such problems can be imagined as a mountain range in which the higher elevations correspond to configurations that violate most of the criteria — enemies sitting next to enemies, so to speak — and the lowest points correspond to solutions in which most or all of the criteria are satisfied. The trick is to find those low points. A conventional adiabatic computer can do that through the equivalent of huffing and puffing over the mountain passes, systematically looking for dips. But a quantum adiabatic computer does a rapid global search. It starts with the analogue of tipping water onto a flat landscape — a state in which the qubits are in a perfect quantum superposition of zeroes and ones — then lets the mountains rise slowly, so that the water naturally pools in the best solutions.

The key to such a computer is that its qubits are meant to stay in their lowest energy state at all times — the precariously balanced pencils have already fallen over. This gives it the massive advantage of being relatively resistant to outside interference, so that little or no error correction is needed until the computer has thousands of qubits or more. And although it is not very useful for factorizing large numbers — the thing that spurred research into quantum computers in the first place — its approach could potentially be used on applications ranging from language translation and voice recognition to working out flight plans for spacecraft.

In 2003, little was known about how to make or program an adiabatic quantum computer, and no one had put in the money and time to build a prototype. Rose decided that D-Wave should try.

Using qubits made from superconducting loops of niobium, cooled to 20 millikelvin above absolute zero to keep them in their lowest energy states, D-Wave's engineers created a usable computer before even they were sure how it worked. "The name of the game from the outset was to make a functional computer," says Williams. "Then they could probe it to see where it was operating correctly."

From there, D-Wave ramped up quickly. The company's 2007 demonstration used a 16-qubit device. By 2011, the D-Wave One machine purchased by Lockheed Martin had 128 qubits (see *Nature* **474,** 18; 2011). This year's D-Wave Two, the model acquired by Google and collaborators including NASA, has 512 (see *Nature* http://doi.org/mt2; 2013). Their computer looks like the proverbial black box: it is a shiny black cube about the size of a sauna. Most of the space is occupied by a cryogenic cooling system; the quantum chip itself is the size of a fingernail. D-Wave aims to double the number of qubits on that chip every year.

HOSTILE AUDIENCE

From the start, D-Wave generated a lot of bad feeling. "I think it is not too strong to say they were initially ridiculed by the academic community," says Jeremy O'Brien, a physicist at the University of Bristol, UK, who invented the computer that can factorize 21.

The problem was not so much the adiabatic-computing approach — it has a solid, if sparse, academic history — but the company's brash style. Most quantum-computing experts feel that Rose and his colleagues should have started by soberly publishing papers characterizing their qubits, rather than putting out press releases. Scott Aaronson, a computer scientist at the Massachusetts Institute of Technology in

Cambridge and a long-time D-Wave sceptic, remains unimpressed by what the company has actually shown that it can do. "They are marketing types who are trying to make the most dramatic claims possible," he bristles.

Rose neither denies nor apologizes for the brashness. He has frequently been quoted as saying, in effect, that his approach is how you build a company. Rose also insists that he has no regrets about the company's 2007 press event — particularly given that it got the attention of Google, which started working informally with D-Wave soon afterwards. "We're not in this business to be popular," he says.

Business style aside, the D-Wave computer is so different from anything else that exists that not even experts know exactly how to judge it. "You do these demonstrations, and how do you know if it's any more significant than factoring 15?" says John Martinis, a physicist at the University of California, Santa Barbara, who heads one of the leading groups working on gate-model quantum computers.

Some of the suspicion is easing as it becomes clearer how the computers operate. In 2011, D-Wave published evidence for quantum behaviour in its 8-qubit chip[3]. Outside the company, the group that has spent the most time on the question is the University of Southern California's Quantum Computing Center in Los Angeles, set up in collaboration with Lockheed Martin when the firm bought its D-Wave computer. In April, a team including the centre's scientific director, Daniel Lidar, circulated results seeming to confirm that the 128-qubit D-Wave One works on a quantum level[4] — although in the fuzzy quantum world nothing is certain, and the results have been challenged[5,6].

Still, D-Wave has chipped away at its credibility problem, concludes O'Brien, "and now they're taken ever more seriously".

ARLEN REDEKOP/VANCOUVER SUN

Geordie Rose expects his company's quantum machine to change the face of computing.

PRACTICAL CONSIDERATIONS

Regardless of how the D-Wave computer works, the practical question is whether it can be used for real-world problems. It can — sort of. In 2009, for example, a Google research team developed a D-Wave algorithm[7] that could learn to judge whether or not a photo showed a car — an example of a 'binary image classifier' that could in principle be used to tell whether a medical image shows a tumour, or a security scan shows a bomb. Finding ever-better ways of doing this sort of task is at the heart of artificial intelligence, and is one area in which an adiabatic quantum computer is expected to excel.

In 2012, researchers at Harvard University in Cambridge, Massachusetts, used a D-Wave machine to find the lowest-energy folding configuration for a protein with six amino acids[8]. They did not have enough qubits to code the problem properly, but even so, on a problem that no other quantum computer could touch, the D-Wave machine found the best solution 13 times out of 10,000 runs. And many of the other answers were good solutions, if not the best.

Meanwhile, Lockheed Martin researchers have developed an algorithm that allows D-Wave machines to tell whether a piece of software code is bug-free[9] — something that, they note, is impossible with classical computers. "You would never know" for sure if a piece of classical-computer code was clean, says Ray Johnson, chief technology officer for Lockheed Martin in Bethesda, Maryland. All anyone could say was that no fault had been found after years of testing. "But now you can say with certainty," says Johnson. "We have great hope, and confidence, in the ability of the computer to scale to real-world complex problems."

D-Wave also competes well against conventional computers in terms of speed, although direct comparisons are difficult. Earlier this year, D-Wave asked Catherine McGeoch, a computer scientist at Amherst College in Massachusetts, to put the D-Wave Two through its paces to satisfy Google before the Internet giant confirmed its deal. McGeoch found that in the optimization-type problems that the D-Wave was designed to solve, it came up with the right answers in half a second, compared with 30 minutes for a top-level IBM machine[10]. "That's one of the most exciting things to happen in quantum computing," says O'Brien.

It is far from clear how long that advantage will last, however, if only because there is no good theory to describe how quantum adiabatic computers will behave on a larger scale. "We are absolutely certain we can build the next generation of this device, but we have absolutely no idea how well it will work," laughs Rose. And since McGeoch presented her results[10] at a meeting in May, other computer scientists have been trying to write yet-faster codes for classical computers. Aaronson says that speed should not be taken as proof of how the device is working. "Even if the machine does get to a solution faster than an ordinary laptop," he says, "then you still face the question of whether that's because of quantum effects, or because a team of people spent $100 million designing a special machine optimized to these types of problems."

In the meantime, work continues to make qubits for universal gate-model quantum computers more reliable, or easier to mass-produce. O'Brien, who admits that his 4-year-old daughter can factorize 21 faster than his computer, is optimistic about the future. "In 10 years' time, I'd be hugely disappointed if we didn't have a machine capable of factoring a 1,000-bit number, involving millions of qubits," he says.

But Rose remains a devotee of the adiabatic church — and is convinced that D-Wave's next generation will prove that it can solve exponentially more difficult problems without taking exponentially more time. "There's going to be absolutely no hope for classical computers if this thing next year behaves as we expect," he says. Rose goes so far as to consider the hardware problem solved: the real challenge, he says, will be the software. "Programming this thing is ridiculously hard," he admits; it can take months to work out how to phrase a problem so that the computer can understand it. But D-Wave has teams working on that — including Rose.

Rose expects tough competition. But with his instinct for fighting, he seems ready for it. ■

Nicola Jones *is a freelance writer near Vancouver, Canada.*

1. Martín-López, E. *et al. Nature Photon.* **6,** 773–776 (2012).
2. Farhi, E. *et al. Science* **292,** 472–475 (2001).
3. Johnson, M. W. *et al. Nature* **473,** 194–198 (2011).
4. Boixo, S. *et al.* Preprint at http://arxiv.org/abs/1304.4595 (2013).
5. Smolin, J. A. & Smith, G. Preprint at http://arxiv.org/abs/1305.4904 (2013).
6. Wang, L. *et al.* Preprint at http://arxiv.org/abs/1305.5837 (2013).
7. Neven, H., Denchev, V. S., Rose, G. & Macready, W. G. Preprint at http://arxiv.org/abs.0912.0779 (2009).
8. Perdomo-Ortiz, A., Dickson, N., Drew-Brook, M., Rose, G. & Aspuru-Guzik, A. *Sci. Rep.* **2,** 571 (2012).
9. Pudenz, K. L. & Lidar, D. A. *Quantum Inform. Process.* **12,** 2027–2070 (2013).
10. McGeoch, C. G. & Wang, C. *Proc. ACM Intl Conf. Comput. Front.* http://dx.doi.org/10.1145/2482767.2482797 (2013).

QUAKE CATCHER

WITH EARTHQUAKE DEATH TOLLS RISING, ROSS STEIN IS BUILDING A GLOBAL RISK MODEL TO MITIGATE FUTURE DISASTERS.

BY JOANNE BAKER

JAMES BROWN/FROM J TO B

In a darkened room in Pavia, Italy, a jumble of stubby arrows spreads out across a large screen like a swarm of ants on the march. To Ross Stein, the marks on this map of the Balkans reveal where earthquakes are most likely to strike, and he urgently wants to share what he sees.

Stein, a geophysicist with the US Geological Survey (USGS) in Menlo Park, California, jumps up from his chair and runs his hand in an arc down the map. Seated in the room are eight seismologists from the former Yugoslav republics and Albania who are analysing their data together for the first time. Ross explains to them how compression is thrusting rocks upwards along faults in some areas and pushing them sideways in others. That pent up energy could be released in devastating tremors, he says, just as it was in July 1963 in Skopje, Macedonia, killing more than 1,000 people.

Such a comprehensive view of the quake risks in the Balkans has been missing, in part because researchers there have limited funding and because some nations prefer to sell geological data rather than disseminate it for free. Two of the workshop's participants, from Slovenia and Albania, are long-time collaborators who could not afford to meet face-to-face in the past decade.

Stein aims to change all that — in the Balkans and elsewhere — by bringing people and data together. He is one of the leaders of the Global Earthquake Model (GEM), an ambitious project to build an open-source digital network of databases and tools focused on seismic dangers around the world. By helping nations, businesses and researchers to assess and minimize risks, Stein hopes to counter the conditions that have led earthquake death tolls to rise over the past century as cities — many with poor building practices — have swelled in quake-prone regions.

After more than five years in development, GEM is nearing major milestones. Next week, the project will release a database of quakes that have occurred over the past millennium, along with a basic version of its software engine, OpenQuake, which will allow users worldwide to calculate their vulnerability to seismic shocks. In December, GEM will unveil a list of all known active faults in the world.

"You'd think that our community would have an inventory, but no one's tried to build one," Stein says. "That's what GEM plans to do."

Over the course of 2014, GEM will add in information about buildings and socio-economic indicators, such as poverty, which could help cities such as Istanbul in Turkey decide how to prioritize the strengthening of vulnerable schools and hospitals.

"It's extraordinary to me how much they have accomplished," says Lori Peek, a sociologist and co-director of the Center for Disaster and Risk Analysis at Colorado State University in Fort Collins, whose research has informed the project.

Leading the GEM effort has marked a major career shift for Stein, a well-respected researcher who has frequently appeared in the media warning citizens about quake risks in the United States. Now he is on a much bigger stage, trying to drum up support for this international project from scientists, governments and companies. "It's been quite an education," he admits.

And it is far from over. Stein must still complete GEM and demonstrate its value. Some critics charge that the effort will not save many lives by offering more sophisticated assessments of seismic risk. Roger Bilham of the University of Colorado at Boulder says that corruption, ignorance and poverty are much greater barriers to safety than lack of information about quakes.

STRESSFUL START

Stein, 59, got his first big taste of seismology as a teenager in Los Angeles, when "terra firma became jello" during the 1971 San Fernando Valley quake, which killed 65 people. But he did not settle on studying Earth science until his college room-mate at Brown University in Providence, Rhode Island, introduced him to the joys of field trips. Stein started a doctorate in glaciology at Stanford University in California and endured the "coldest, wettest, windiest fieldwork". Then, wanting to pursue a topic with social impact, he switched to earthquakes after hearing a talk from a USGS scientist. He joined the agency in 1981.

In his research, Stein has focused on how an earthquake in one spot transfers stress to other regions. His modelling efforts have provided a means of estimating whether tremors will increase or decrease the likelihood of earthquakes elsewhere.

That and other work, notably in Turkey and Japan, made Stein the second most highly cited earthquake scientist from 1993 to 2003. And his impact has spread far beyond the research community. He has appeared in numerous documentaries and is often in front of a camera after a large quake.

Stein's research trajectory was drastically altered by the 2004 Sumatra–Andaman earthquake and tsunami, which killed more than 230,000 people in 14 countries. That event, he says, "crystallized our failure as a community" by revealing how little scientists had done to help the region to prepare for the hazards expected in that area. "In some ways I felt there was blood on my hands," says Stein.

He decided that it was more important to address seismic risks in poor countries than in California or Japan, where a long tradition of research and strong building codes has already reduced dangers. From Jakarta to Port-au-Prince, urban populations are skyrocketing near major faults and along tectonic-plate boundaries. The influx of people is filling poorly constructed houses that become death traps in quakes, Stein says. Seismologists predict that, before long, a large shock will kill a million people.

In 2006, after an earthquake workshop in Potsdam, Germany, Stein and two other seismic-risk experts — Jochen Zschau at the Helmholtz Centre in Potsdam and Domenico Giardini of the Swiss Federal Institute of Technology (ETH) Zurich — decided to combat that trend by setting up GEM.

A raft of international, governmental and non-governmental organizations already helps at-risk communities to prepare for and respond to quakes, but those efforts are fragmented. The Office of US Foreign Disaster Assistance (OFDA), which sponsored Stein's Balkan workshop, helped to develop a tsunami warning system in Indonesia after the 2004 event and is running seismic projects in Haiti and the Dominican Republic. And GeoHazards International (GHI), a non-profit organization based in Menlo Park, has worked in more than 20 countries to raise awareness and train construction engineers in quake safety. But no single organization can span every town and city, and no country can afford to reinforce or insure every building. Knowing where risk is highest is key, Stein says.

Information is also splintered. Peek, who advises the GHI, participated in a study for the GEM consortium that showed that communities from San Francisco, California, to Chincha in Peru all need a central resource on earthquake risks — one that pools data on seismic threats, construction issues, and economic and social factors. That would help local officials to prioritize which public buildings or regions to strengthen, and allow emerging cities such as Kathmandhu or Lima to plan how to grow without increasing their seismic risk.

GEM aims to provide that resource through OpenQuake. Built using a geographical information system, this platform will include analytical tools that allow anyone — scientists, governments and companies — to estimate the chances of economic and human losses from earthquakes (see 'Trouble spots').

The calculators will draw on the GEM's global databases of quakes, faults, housing types and socio-economic information, which are being rolled out over the next 18 months. In January, GEM released a reference catalogue of more than 20,000 global earthquakes of magnitude 5.5 and above that have occurred since 1900. To produce it, the consortium reanalysed all the seismic data involved, improving estimates of earthquake epicentres and magnitudes. It is the biggest resource of its type and has allowed seismologists to see, among other things, how seismic activity concentrates on a major fault below Guatemala, says Stein.

To get the project off the ground, Stein and his collaborators had to persuade funders to back the plan. The Paris-based Organisation for Economic Co-operation and Development (OECD), which sponsored the Potsdam workshop where the GEM idea was seeded, gave Stein and the founders access to governments officials. In the wake of the Sumatran tsunami and a major quake in Pakistan, OECD member states in high-risk regions wanted to minimize their exposure to giant economic losses.

"WE WILL FACE ABUSE. SOME GOVERNMENTS WILL PUSH BACK AGAINST GEM'S ASSESSMENTS."

Stein's contacts grew from there. In 2007, Munich Re became the first company to get involved, giving €5 million (US$6.6 million) over five years. It saw an opportunity in the global data being collected by GEM, which could help insurance companies and reinsurance brokers to diversify their portfolio to avoid being wiped out by a single earthquake.

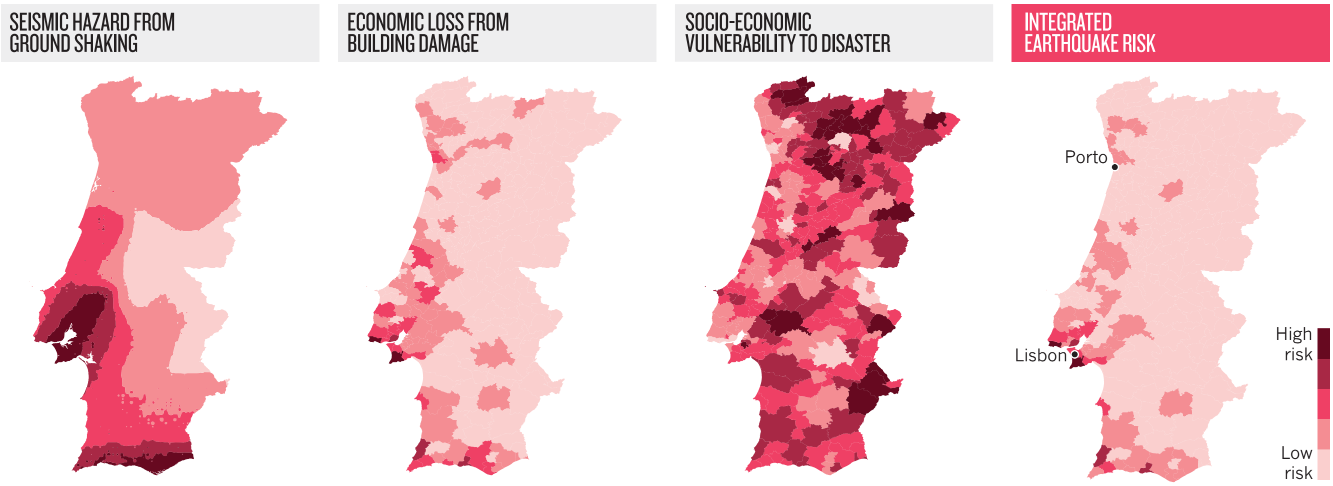
Today, 16 governmental agencies, such as the OFDA, and 10 insurance and engineering companies have joined GEM, which is a non-profit public–private partnership headquartered in Pavia and has some 20 staff. These sponsors have contributed more than 90% of the €24 million needed to release the full OpenQuake platform, which is planned for November 2014. In addition, nine organizations, including the World Bank, have become associate non-paying members.

Dealing with the disparate interests has been a steep learning curve for Stein. "All have a stake," he says, and "issues to champion". At GEM board meetings, he says, the different sectors sit in groups around a U-shaped table — the countries on one arm, companies on the other and the non-governmental organizations in the centre.

Some scientists, however, are unhappy that companies have a seat at the table at all. "Suppose you could manipulate hazard forecasts to

TROUBLE SPOTS

The Global Earthquake Model has tools to assess earthquake risk by combining data on ground shaking, construction practices and socio-economic vulnerability. An example from Portugal shows the integrated risk from a magnitude-8 earthquake such as the one that destroyed Lisbon in 1755.



SEISMIC HAZARD FROM GROUND SHAKING

ECONOMIC LOSS FROM BUILDING DAMAGE

SOCIO-ECONOMIC VULNERABILITY TO DISASTER

INTEGRATED EARTHQUAKE RISK

SOURCE: GEM

justify higher quake-insurance premiums in built-up areas," muses Robert Geller at the University of Tokyo. But Stein is pragmatic. "If you are talking to a finance minister you have to talk about economics or they won't pay attention," he says.

To be widely used and trusted, Stein says that GEM must be seen as independent, transparent and accessible. That's why OpenQuake uses open-source software — and why GEM plans to give away the project's data and products to anyone, including the public, scientists and governments, if they are engaged in non-commercial work. Companies wanting to use the data commercially will need to sponsor the organization. Governmental agencies are asked to make a contribution that is proportional to their total investment in research and development. For Ecuador, that runs to €15,000 per year, whereas Germany is paying €275,000 annually.

The founders hope that banks and companies will join in order to build new markets or products. They could use GEM data and tools to develop 'catastrophe bonds', a type of insurance in which investors take the risk in return for payments if a specified event does not occur. Companies have offered such bonds since the mid-1990s, but governments are now getting in on the act. Earlier this year, a group including the Turkish government issued a bond that will release US$400 million if Istanbul experiences a major shock in the next three years.

SCIENCE DIPLOMACY

Stein, who chairs GEM's scientific advisory board, has to do more than marshal the seismic data and models. He is part of the human glue that melds the sectors together, a post that requires the skills of a diplomat and a salesman.

Both skill sets were on display at the Balkan workshop, where the assembled seismologists began arguing over funding inequities and other problems in previous regional initiatives to analyse earthquake risk. At one point, some participants shouted at each other across the table. Stein let them have their say and then stepped in to calm the waters. He asked each in turn to express their views and offered to visit each country that autumn, to convince government officials and university heads to back the researchers.

As GEM becomes more visible, Stein knows that he will have to contend with critics. Some members of the seismology community say that it is misleading to map hazards on the basis of past earthquakes because the historical record is too short, and large earthquakes often occur where none has previously been witnessed. In northeastern Japan, for example, risk maps for the Tohoku region did not anticipate a monster quake of the size that struck in 2011. Other researchers, such as Bilham, question whether the project's engineering goals will ever be enacted; they argue that many countries already have adequate building codes but fail to enforce them — so better risk models won't help.

Stein has dealt with some of the criticism by inviting naysayers to participate in GEM. Seth Stein (no relation), a seismologist at Northwestern University in Evanston, Illinois, who is a long-standing opponent of some seismic-hazard maps (see *Nature* **479,** 166–170; 2011), attended a GEM workshop. Although Seth Stein sees GEM's open-source, standardized and modular approach as "a good step in the right direction", he also hopes that the seismology community will take advantage of the resource to do broader analyses exploring the limitations of seismic-hazard analysis.

Looking forward, Ross Stein seems most concerned about securing funding. GEM will need more subscribers to pay for the curation and updating of its databases in the future and is seeking a further €10 million to fund allied regional programmes to enhance the local detail of the risk databases. To attract and retain sponsors over the long term, the project must keep rolling out useful features on related risks — such as models including tsunamis, landslides and liquefaction, which happens when seismic shaking weakens soil to a point at which it begins to behave like a fluid.

The most difficult challenge long term, however, may be handling the backlash over risks identified by GEM. Stein says that GEM "is not an advocacy organization" and will not get involved in policy decisions on the basis of its assessment. Even so, "we will face abuse", Stein accepts. "Some governments will push back against GEM's assessments because they differ from their priorities."

In Pavia, as the Balkan workshop winds up, Stein practises the diplomatic skills he will need to make GEM succeed. Moving beyond the earlier rancorous discussion, he suggests that all the participants write a joint publication and apply to the OFDA for funds to enable them to meet again in six months. All the seismologists pledge to continue the collaboration. Such a meeting will be essential "if we want to build a harmonized model for the whole Balkan area", says Barbara Sket Motnikar of the Jožef Stefan Institute in Ljubljana.

Three weeks later, the OFDA agrees to fund a second workshop for the group. The decision underscores some of Stein's parting words to the Balkan seismologists: "Never underestimate the power of your enthusiasm." ■ SEE EDITORIAL P.271

Joanne Baker *is a Comment editor for* Nature *in London.*

TYRONE TURNER/NATL GEOGR. SOC./CORBIS

Residents of Greensburg, Kansas, rebuild their community with energy-efficient homes after a tornado destroyed the town in 2007.

Positive energy

To change attitudes towards energy scarcity and climate change, focus on transitions and solutions, not danger and loss, says **Chris Nelder**.

"I'm sometimes asked if I'm optimistic or pessimistic about energy," economist Daniel Yergin admitted in 1979, when he was a lecturer at Harvard Business School in Cambridge, Massachusetts. Concerned that the United States was doing little to maintain its oil production, which had dropped since 1970, he declared himself a pessimist. He was referring to 'peak oil': the idea that global oil production will peak and then decline. The United States was mired in an energy crisis with no easy way out.

That is still true today, but Yergin is now one of the loudest voices telling world leaders a tale of future oil abundance. In a 2011 editorial[1] in *The Wall Street Journal*, he asserted that advances in technology continue to make new volumes of oil viable, and that the peak "is still not in sight".

Such messages of abundance are the norm in the media and in policy circles. The fact that predictions of rising oil production have been consistently proved wrong does not dim their appeal. Oil pessimists have offered more accurate guidance since peak oil was first predicted in 1956 by M. King Hubbert, a geologist who worked for Shell Oil in Houston, Texas, and for the US Geological Survey. Why has their story gained no traction?

I believe that people simply want to hear a positive message. Too often, scientists and analysts invite indifference and resistance by framing energy and climate-change debates in terms of danger and loss. It does not help that these complex topics can only be understood with a grasp of highly technical definitions and concepts, and mitigated only through arcane policy measures.

Telling an optimistic story, by using the language of solutions, transitions and resilience, is more persuasive and more likely to promote useful action. A small rural ▶

Transition towns such as Groningen in the Netherlands use local initiatives to produce renewable energy and to make their communities less dependent on oil.

B. SCHMID/AMANAIMAGES/CORBIS

town is unlikely to build a wind farm to fight climate change, but it might support such a project if it is seen as a way to create jobs and to improve the local economy, while empowering the community and enhancing its self-reliance.

TALES OF ABUNDANCE

Agencies such as the International Energy Agency (IEA) in Paris and the US Energy Information Administration (EIA) in Washington DC, as well as big banks and the media, are constantly reassuring us that energy supplies will meet future demand, and that technological advances will bring energy at an acceptable price. The boom in US tight oil (from shale formations) is touted as a 'game changer', with the United States purportedly poised to surpass Saudi Arabia as the world's leading oil producer by 2020 (ref. 2).

In reality, it is becoming increasingly difficult to deliver fuel at an reasonable price. Global production of conventional oil stopped growing at the end of 2004. From 2004 to 2012, investment by the oil industry doubled to US$600 billion a year, and oil prices nearly quadrupled. But average annual production increased by only 4.3%. This has acted as a brake on the global economy.

The EIA's "highly uncertain" estimate for unproved, technically recoverable (but not necessarily economically viable) tight-oil resources in the United States is 58 billion barrels — enough for only 8.6 years' worth of US consumption[3]. And it is commonly asserted by the media and the energy industry that the United States has a 100-year supply of gas, but proven dry-gas reserves will last only 12.5 years at the 2012 rate of US consumption.

Even the American Petroleum Institute in Washington DC, the US oil industry's main lobbying group, admits that world oil supplies "have been struggling to keep up with rising demand". In 2010, the EIA's review of its own forecasting history found that it had badly underestimated the prices for oil, natural gas and coal for more than a decade[4].

The IEA anticipated that Middle Eastern oil supply would double between 2000 and 2030, with another 10% added by Canadian tar sands, heavy oil from Venezuela and gas-to-liquids, a process in which natural gas is converted to liquid fuels. And a few years ago, biofuels were seen as a solution. These predictions now seem absurd.

Ethanol turned out to be an expensive way to make low-quality fuel, driving up food prices and sparking 'tortilla riots' in Mexico in 2007. Oil production from Canadian tar sands reached 1.6 million barrels a day in 2012, just over half of what was projected in 2006. Heavy oil and liquid fuels processed from coal have yet to scale up affordably. The 'hydrogen economy', touted in 2005 by policy-makers and industry representatives as a transformational vision, faded without an epitaph.

Such abundance stories are generally based on econometric models that chart a path to economic growth. They are not based on actual resources and make assumptions that may be biased. Many are flawed (see go.nature.com/3bjdu3).

Scientists and economists with decades of experience in oil-and-gas companies and energy agencies expect the global supply of liquid fuels to start declining before 2020. These forecasters seem closer to the mark, and include Jean Laherrère of Total in Paris, BP geologist Colin Campbell, geologist and social entrepreneur Jeremy Leggett, former IEA oil analyst Olivier Rech, and economist Michael Kumhof of the International Monetary Fund in Washington DC. Yet few people have heard of them, and the media generally disregard them. Meaningful public debate over energy policy has been stifled in the process[5].

CHANGING BEHAVIOUR

Scientists must to learn how to tell as compelling a story about energy, climate change and resource scarcity as advertisers or lobbyists. For a person to relate to a story, it must be consistent with what their community believes[6]. As psychologist Dan Kahan has noted: "People endorse whichever position reinforces their connection to others with whom they share important commitments."[7] This explains, for example, why groups at either end of the political spectrum can hold identical views on issues as disparate as same-sex marriage and climate change.

We must also adapt our communications to allow for the fact that most thinking is automatic and does not follow rational logic, as the behavioural scientist and Nobel laureate in economics Daniel Kahneman points out. We trust narratives that fit our emotions, associations and experiences, rather than actively assessing the evidence. This is why the peak-oil story gained currency in the press in 2008, when prices for oil and gasoline shot up — it fitted in with our experiences. When prices fell, the story faded. A Google Trends query of news headlines displays a strong relationship between the search terms 'peak oil', 'oil prices' and 'gasoline prices' (see go.nature.com/p3ihnm).

Similarly, extreme weather events such as hurricanes and tornadoes capture the public's attention in a way that decades of warnings about global warming have failed to do — notwithstanding that the connection between the two is complex. Hurricane Sandy, which devastated the east coast of the United States in October 2012, emboldened the governor of New York, Andrew Cuomo, to declare that "anyone who says there hasn't been a dramatic change in weather patterns is in denial". A Google search finds more than one million results that mention both 'Hurricane Sandy' and 'renewable energy'.

A story must also be positive to be amplified in the press. Accuracy has become boring in the world of 'link-bait' journalism: editors and journalists want to publish stories that are popular. If we want action on energy transition and combating climate change, we must offer concrete and viable solutions — no money goes to problems, only to fixes. We should advocate solutions in an upbeat, tractable way, tailored to particular world views.

TALES OF TRANSITION

R. Rex Parris, the mayor of the Mojave Desert town of Lancaster, California — a right-leaning, suburban, middle-class community where many people work at an Air Force base — has taken that positive tack in his push to make his city the first in the United States to produce zero net carbon emissions.

"We can't fix [climate change] top-down, but it's easy to fix bottom-up," Parris said at an energy conference in April. Instead of scaring his citizens, he has leveraged the authority of the city's building and planning departments to encourage solar power. Lancaster now has the most solar energy production per capita of any city in California.

It has also become the first city in the state to require that developers of new homes build at least 1 kilowatt of solar capacity for every home they construct. The small California town of Sebastopol has followed suit, requiring solar photovoltaic systems on all new buildings, major additions and remodellings.

This bottom-up approach has also worked for the small farming town of Greensburg, Kansas. After a tornado destroyed most of the town in 2007, the residents came together and "right off the bat, they started talking about green buildings", said the town's mayor, Bob Dixson. They crafted a new mission statement that focused on "working together for future generations", which required all new buildings to be certified by the Leadership in Energy and Environmental Design ratings system for sustainable buildings. The town's new community wind farm, built through a public–private partnership, exports some of its 12.5 megawatts of power to other nearby towns.

Founded in the United Kingdom, the Transition Town movement is a grassroots network of communities organized in 2005 in response to the problems of peak oil, climate destruction and economic instability. I would argue that this initiative has done more to build resilience to these threats than peak-oil modellers ever did. Transition towns, which include places as diverse as Tucson in Arizona and Groningen in the Netherlands, find ways to make their localities more sustainable, by creating community gardens, building solar-power systems and staging river clean-up events, among other things. The network now includes thousands of towns across the globe.

Why have these local approaches worked? People like feeling that they are part of the solution, instead of being hostage to intangible problems such as oil dependency and climate change. They seize on things that give them hope and optimism. Capitalizing on these feelings should be a common objective of those who seek sustainable solutions.

"Local measures must be championed to move us away from dependency on fossil fuels."

I believe that these local measures must be championed to move us away from dependency on fossil fuels and towards renewable energies, and away from personal vehicles to public transport. Instead of agitating for indirect and punitive policy mechanisms such as carbon taxes, we should be advocating feed-in tariff schemes that encourage renewable energy, better rail networks, bicycle-friendly streets, local-food production, and improvements to the efficiency of our built environment (see go.nature.com/jly7qv).

Increasing railway usage could reduce oil demand permanently and at scale. But we should highlight its lifestyle virtues: trains are a safer, cheaper, more relaxing and more productive mode of transport. Since the oil-price spike of 2008, several studies have shown that commuters prefer using public transport for these reasons.

Messages about climate change and energy that use fear- and threat-based tactics have not mobilized responses. Let's try a positive tack instead. As the old sales saying goes: "Sell the sizzle, not the steak." ■

Chris Nelder *is an energy analyst and consultant based in Marin County, California. He blogs at GetRealList.com.*
e-mail: chris@getreallist.com

BEN MARGOT/AP

A resident of Sebastopol, California, adjusts a solar panel. The town requires photovoltaics on new homes.

1. Yergin, D. 'There will be oil' *The Wall Street Journal* (17 September 2011).
2. International Energy Agency. *World Energy Outlook 2012* (IEA, 2012).
3. US Energy Information Administration. *Technically Recoverable Shale Oil and Shale Gas Resources* (EIA, 2013).
4. US Energy Information Administration. *Annual Energy Outlook Retrospective Review: Evaluation of Projections in Past Editions (1982–2009)* (EIA, 2010); available at http://go.nature.com/x1yett.
5. Littlefield, S. R. *Energy Policy* **52,** 779–788 (2013).
6. American Psychological Association. *Psychology and Global Climate Change* (APA, 2009); available at http://go.nature.com/54yxp3.
7. Kahan, D. *Nature* **463,** 296–297 (2010).

Honeybee (*Apis mellifera*) workers busy themselves on their comb.

ENTOMOLOGY

The apian way

Mark Winston revels in a deep exploration of the honeybee colony and its organization.

In his 1901 book *The Life of the Bee*, philosopher, poet and Symbolist Maurice Maeterlinck mused over the "spirit of the hive". How, he wondered, do many thousands of social insects organize themselves into a cooperating colony? In the fascinating *The Spirit of the Hive*, eminent bee geneticist Robert Page demonstrates how science is answering that key question.

The book — representing a lifetime's research for Page — illustrates just how far we have come in dissecting and reconstructing the myriad factors responsible for colony-level functions. Page stands on the shoulders of, and acknowledges, bee geneticist Harry Laidlaw, who unravelled many basic aspects of honeybee genetics and mating behaviour. But Page chronicles the expansion of methodologies over many decades, from behavioural and hormonal analyses to neurochemistry, and eventually DNA sequencing. We also learn how researchers working with honeybees pioneered many techniques. An example is genetic mapping with quantitative trait loci, which can identify the multiple linked genes that determine complex behaviours such as pollen foraging.

The Spirit of the Hive hinges on the simple question of what causes honeybees to collect more or less pollen and nectar. Individual worker bees were once viewed as behaviourally identical, but Page and others have demonstrated that there is significant variation in individuals' responses to stimuli that determine the level of nectar and pollen foraging, such as the amount of pollen stored in the nest. Furthermore, each queen mates with an average of 12 drones, whose sperm mixes in her sperm sac. Each colony's worker bees thus present a distinctive mix of genetically influenced tendencies for foraging.

Page then tackles how individual bees with differing genetic backgrounds create flexibility in the colony to regulate tasks — including pollen, nectar and water collection, defence and the removal of dead bees from the nest — in response to environmental conditions. This collective effort is one of the best examples we have of the relative significance of genetics, environment, nature and nurture for social organisms. The colony's needs turn out to be the main driving factor in pollen collection, emphasizing the importance of environmental factors. Page says that the bee's genotype explains just 8–25% of behavioural variance, and usually closer to 8%. This is a comforting concept for those who prefer nurture to nature.

The Spirit of the Hive: The Mechanisms of Social Evolution
ROBERT E. PAGE
Harvard University Press: 2013.

The Spirit of the Hive is pitched at Page's fellow evolutionary biologists, and in places requires some understanding of Boolean

SIMON COLMER/NATUREPL.COM

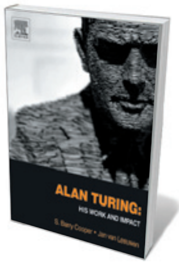
logic, binary behaviour and basic genetics. But the book is at its best when Page uses story and description rather than models and mathematics. He clearly has considerable affection for his co-workers and for bees, and provides marvellous glimpses into how research is conducted. He opens that world to us through descriptions of his own elegant experiments, such as those exploring the genetic component of pollen collection.

I would have liked to hear more about the people he worked with and his relationship with his insects after so many decades in the field. Page finishes abruptly, with a too-short chapter that brings us back to Maeterlinck's *The Life of the Bee*. I missed the soaring language of the poet here, which would have beautifully completed the circle — for instance, Maeterlinck's description of the hive as containing "the enigma of intellect, of destiny, will, aim, means, causes; the incomprehensible organization of the most insignificant act of life".

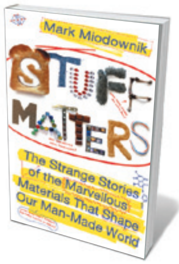
Still, Page's book is a delightful example of how one dedicated career in science can dramatically deepen and broaden our perceptions of the world around us. ■

Mark L. Winston *is academic director and a fellow of Simon Fraser University's Centre for Dialogue in Vancouver, Canada.*
e-mail: winston@sfu.ca

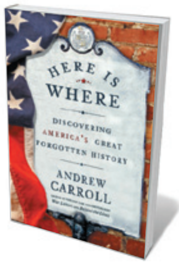
Books in brief



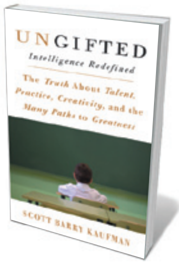
Alan Turing: His Work and Impact
Edited by S. Barry Cooper and Jan van Leeuwen ELSEVIER *(2013)*
The new testament of computer science has come, 101 years after the birth of founding prophet Alan Turing. It took 70 renowned evangelists from all walks of science and philosophy to put the polymath's words in context and dissect his living impact on pure maths, physics, biology, engineering, banking, metaphysics and beyond. How big is the incomputable universe? Can digital machines think? Do daisies emerge from pure chemistry? If your soul craves answers to such questions, this is your new bible.


Mark Miodownik
STUFF MATTERS
The Strange Stories of the Marvellous Materials That Shape Our Man-Made World

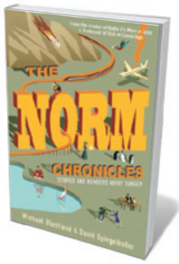
Stuff Matters: The Strange Stories of the Marvellous Materials that Shape Our Man-Made World
Mark Miodownik VIKING *(2013)*
Today's materials vie with our wildest imaginings, from two-dimensional graphene to aerogel — made up of 99.8% air, resembling "solid smoke" and created by NASA to gather space dust. In this homage to materiality, Mark Miodownik tells us why we should care about stuff. The materials specialist traces his obsession back to a violent childhood epiphany when, stabbed with a razor blade, he woke to the wonders of steel. Here, we too are jolted into a new consciousness of the made world's multiple facets.



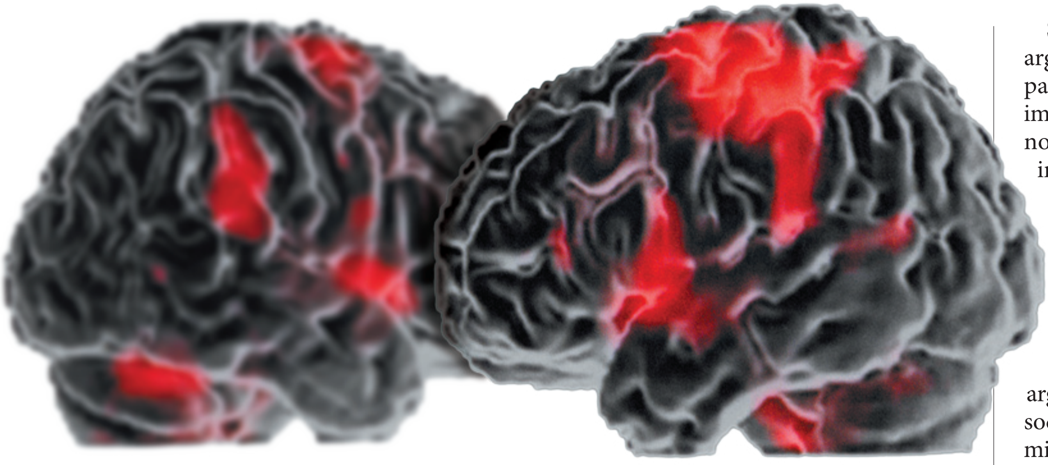
Here Is Where: Discovering America's Great Forgotten History
Andrew Carroll CROWN ARCHETYPE *(2013)*
From Plymouth Colony to Gettysburg, Pennsylvania, hotspots of US history are well and truly mapped. Yet 'off-piste' places with scientific importance abound, as historian Andrew Carroll reveals on this road trip around forgotten America. He ably guides us through triumphs and horrors: the Oregon caves where the continent's oldest human DNA was radiocarbon dated; California's Sonoma Developmental Center, where thousands were sterilized in the name of eugenics; the Massachusetts cherry tree where rocketeer-to-be Robert Goddard dreamed of interplanetary travel; and much more.


UNGIFTED
Intelligence Redefined
The Truth About Talent, Practice, Creativity, and the Many Paths to Greatness
SCOTT BARRY KAUFMAN

Ungifted: Intelligence Redefined
Scott Barry Kaufman BASIC BOOKS *(2013)*
Hearing difficulties and a low IQ score left the young Scott Barry Kaufman labelled learning disabled. Now a cognitive psychologist, he charts his journey from judgement by metrics to a deeper understanding of human intelligence. Gathering research on areas from nature–nurture interplay to the psychology of motivation, he presents a convincing "theory of personal intelligence". But what emerges most clearly is how all children — gifted, disabled or simply humming with untapped abilities — need a fine-tuned, holistic education to shine in their own extraordinary ways.



The Norm Chronicles: Stories and Numbers About Danger
Michael Blastland and David Spiegelhalter PROFILE BOOKS *(2013)*
Writer Michael Blastland and risk specialist David Spiegelhalter offer a fresh take on the hot topic of risk. They explore chance and probability through the characters Prudence, Norm and Kelvin, who represent the spectrum of risk-taking behaviour. Even as the authors offer innovative tools for measuring acute and chronic risk, they remind us that data have limits. If you want to know the odds on shrinking your lifespan by imbibing that second glass of wine or being hit by an asteroid, take a gamble on this book. **Barbara Kiser**



SCOTT T. GRAFTON/VISUALS UNLIMITED

Patterns of brain activity can only tell researchers so much.

NEUROSCIENCE

Rise of the neurocrats

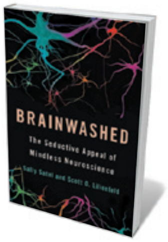
Sandra Aamodt evaluates a cautionary account of how brain-scan results could be used and abused.

Imagine a world run by 'neurocrats' who use brain scans to detect lies, determine why people commit crimes, and control what brand of soap consumers choose. In *Brainwashed: The Seductive Appeal of Mindless Neuroscience*, psychiatrist Sally Satel and clinical psychologist Scott Lilienfeld explain that this vision of brain scanning exceeds our current abilities. They also argue that emphasizing neural causes of behaviour over psychological and environmental ones could undermine our belief in free will and personal responsibility.

Satel and Lilienfeld provide an engaging overview of the technical and conceptual factors that complicate the interpretation of brain scans obtained by functional magnetic resonance imaging and other techniques. One problem is flawed statistical analysis, as illustrated by the Ig Nobel-prizewinning study in which brain 'activity' was recorded in a dead salmon. A deeper concern is that activity in a particular human brain area rarely corresponds to a unique mental state. For example, amygdala activity is often associated with fear, but can indicate surprise, happiness, anger or valuing options.

Sloppy inferences of this sort are common in the popular press and embarrassing for the field, which is why many neuroscientists push back against such errors. One example is the opinion article 'You Love Your iPhone. Literally' (*The New York Times*, 30 September 2011), in which the author wrote that the observation of similar brain activity in response to both a phone and a loved one suggested that they evoke the same mental state. A group of 46 prominent researchers wrote a letter to the newspaper's editor to explain why the conclusions were unconvincing.

Beyond the reputation of neuroscience, misuse of brain imaging has potentially serious consequences for individuals. As Satel and Lilienfeld explain, brain-based lie detectors currently belong in the lab rather than the legal system: both the false-positive and false-negative rates are unacceptable for high-stakes applications. Almost all courts that have evaluated these methods have agreed. A lie detector based on electroencephalography was admitted as evidence in one murder trial in India, but an appeals court ruled the test "unscientific" and ordered that the defendants be released on bail.



Brainwashed: The Seductive Appeal of Mindless Neuroscience
SALLY SATEL AND SCOTT LILIENFELD
Basic Books: 2013.

Lawyers have had more success introducing brain scans during sentencing to support claims of diminished responsibility. Courts have considered evidence of brain damage, incomplete brain maturation in adolescents and unusual brain responses that are consistent with psychopathy. The authors note, however, that even when brain activity differs on average between groups with particular characteristics, the results typically overlap enough that individual scans are not diagnostic.

Satel and Lilienfeld make a philosophical argument too: they believe that moral culpability, blame and even retribution serve important societal functions that we should not relinquish. The authors are responding to suggestions that we should rethink traditional ideas of blame, in light of the growing scientific consensus that many influences on actions and decisions occur outside conscious awareness.

Neuroscientists including David Eagleman, in *Incognito: The Secret Lives of the Brain* (Pantheon, 2011), argue that we should instead attribute antisocial behaviour to brain malfunctions that might be correctable. In one example, a previously law-abiding man became interested in child pornography and tried to molest his stepdaughter. He was found to have a brain tumour and his impulses disappeared when it was removed, but returned when it grew back. In such a case, cancer treatment would protect society more effectively than blame or punishment.

However, Satel and Lilienfeld argue that viewing responsibility from a biological perspective may backfire. Attempts to use neuroscience research to destigmatize mental disorders have produced mixed results. People who believe that conditions such as schizophrenia are strictly biological are less likely to blame people with the condition for their behaviour, but also view them as more dangerous and less able to change.

In addition, this outlook may have adverse consequences for drug addicts. Several population studies have found that more than 75% of addicts eventually quit, many without treatment. The authors speculate that people who think of addiction as a chronic, relapsing brain disease may be less likely to succeed in kicking the habit.

This philosophical debate over the relevance of neuroscience to free will may be less important than it seems because it is difficult to change people's attitudes. New information tends to flow into well-worn cultural paths as people process it under their own assumptions and agendas. Psychologists find that non-specialists are not yet convinced that the self is a construct based in brain activity, nor that biology trumps free will. Instead, people assimilate scientific concepts into their previous ideas about the importance of responsibility and self-control.

In short, the neurocrats are not coming for your thoughts any time soon. In the meantime, *Brainwashed* offers much to bolster popular understanding of what brain imaging can and cannot achieve. ■

Sandra Aamodt *is a former editor of* Nature Neuroscience *and author of* Welcome to Your Child's Brain: How the Mind Grows from Conception to College.
e-mail: sandra.aamodt@gmail.com

Correspondence

Use oil wealth to save Brazil's biodiversity

The production of recently discovered offshore oil in Brazil has reached 300,000 barrels a day, and is expected to rise to 2.1 million barrels a day by 2020. Since the oil was found under a 2-kilometre layer of salt beneath the sea bed in 2006 (see *Nature* **455,** 438–439; 2008), coastal ports have proliferated to keep pace with the boom. We suggest that some of this wealth should go into evaluating the environmental costs of such rapid development, which would help to safeguard the region's rich biodiversity.

Some ports are being constructed in conservation sites, including mangrove swamps and coastal shrub forests called *restingas*. Offshore oil spills are frequent (see, for example, go.nature.com/jshpzd), and pipelines are being deployed in protected areas, such as Brazil's Atlantic Forest.

President Dilma Rousseff has proposed earmarking oil royalties for investment in basic education. However, scientific research will not benefit (see *Nature* http://doi.org/d8zgdk; 2011) — even though it too could help to educate people in mitigating the environmental damage caused by oil production and distribution.

Renan de França Souza *State University of Rio de Janeiro, Brazil.*
renan1604@hotmail.com
Roberto Leonan Morim Novaes, Saulo Felix *Federal University of the State of Rio de Janeiro, Brazil.*

Tapping into success and collaboration

Jonathan Adams' analysis cannot distinguish whether scientists collaborate internationally because they are successful, or whether they are successful because they collaborate internationally (*Nature* **497,** 557–560; 2013).

If it is the former, then any government initiative to promote collaboration for its own sake risks simply degrading the correlation between collaboration and success, rather than improving the quality of scientific output. Impact factors and publication rates are arguably imperfect estimators of past success, and insecure guides to future success.

Michael Weale *King's College London, UK.*
michael.weale@kcl.ac.uk

Social change vital to sustainability goals

David Griggs and colleagues argue convincingly that sustainable development goals (SDGs) should enhance the role of natural capital and ecosystem services within a framework of economic development and poverty reduction (*Nature* **495,** 305–307; 2013). We believe that it is also crucial to factor social change into the SDG process.

It will be essential to motivate, guide and support social change towards sustainable practices at all scales of governance — globally, nationally and individually. Simply setting ambitious goals will not generate these changes: their formulation must include details of the processes needed to achieve them. For example, SDG targets should take into account ideologies, religious beliefs and institutions, including formal and informal rules and customs.

Setting targets in terms of specific and simple changes would help to overcome institutional inertia, induce desirable shifts in governance and lead to changes in people's behaviour. One such example might be to build formal accounting of carbon dioxide emissions and reporting practices into global trade and retail chains as part of international climate agreements.

Another suggestion would be to use the SDGs to create global networks for problem-solving, or social-innovation 'labs', which could catalyse new sets of rules, ways of thinking and processes for action and decision-making.

Albert V. Norström* *Stockholm Resilience Centre, Stockholm University, Sweden.*
albert.norstrom@stockholmresilience.su.se
**On behalf of 17 co-authors. See go.nature.com/i7bjjc for full list.*

Public engagement should start early

Public scepticism about science, compounded by poor communication, is standing in the way of implementing sustainable technologies that could solve pressing global issues, such as the provision of sufficient clean energy, water and food (see go.nature.com/6rs2ih and go.nature.com/yvzaht). I suggest that this obstruction could be alleviated by providing more effective training for young scientists in the advantages of public engagement.

Early-career scientists may feel neither encouraged nor equipped to communicate beyond the scientific community. Current trends in performance-evaluation criteria do not seem to motivate scientists to engage more effectively with society (see go.nature.com/t43rhg).

Postgraduate education needs to include exposure to and training in social responsibility, public communication and leadership. Basic personal engagement with the United Nations Millennium Development Goals, policy-makers, industries and communication specialists, for example, would lead to a more connected cohort of scientists, who could then pass their skills on to the next generation of researchers.

Bernard Slippers *University of Pretoria, South Africa.*
bernard.slippers@fabi.up.ac.za

Priming-effect author responds

I wish to clarify your perspective on David Shanks's failure to replicate our 'intelligence priming' results (see *Nature* **497,** 16; 2013; and clarification at go.nature.com/8ep3nc).

Variations and extensions of the effect of intelligence- or stupidity-related primes on intellectual performance have been widely studied since we published our original paper (A. Dijksterhuis and A. van Knippenberg *J. Pers. Soc. Psychol.* **74,** 865–877; 1998). These so far amount to 27 independent experiments from 12 different labs in eight countries (for instance, see A. D. Galinsky *et al. J. Pers. Soc. Psychol.* **95,** 404–419; 2008; further references available from the author). To my knowledge, Shanks *et al.* are the first to publish a failure to replicate our findings (although see also www.psychfiledrawer.org, which lists two single-study unsuccessful replication attempts and one successful attempt).

Given that the percentage of successful replications in psychology is low (it seems to be 30–40%), there is no reason to assume that the intelligence-priming effect is especially difficult to reproduce. Also, since 1998 we have gained a better understanding of the mechanism underlying the priming effect (see, for example, S. L. Bengtsson *et al. Soc. Cogn. Affect. Neurosci.* **6,** 417–425; 2011).

There are technical and methodological factors in the experimental design used by Shanks and his team that could explain their inability to replicate our results (details available from the author). For example, their general-knowledge questionnaire was unusually difficult and might have itself moderated the effect — an idea that awaits further research.

Ap Dijksterhuis *Radboud University Nijmegen, the Netherlands.*
a.dijksterhuis@psych.ru.nl

Christian de Duve

(1917–2013)

Biologist who won a Nobel prize for insights into cell structure.

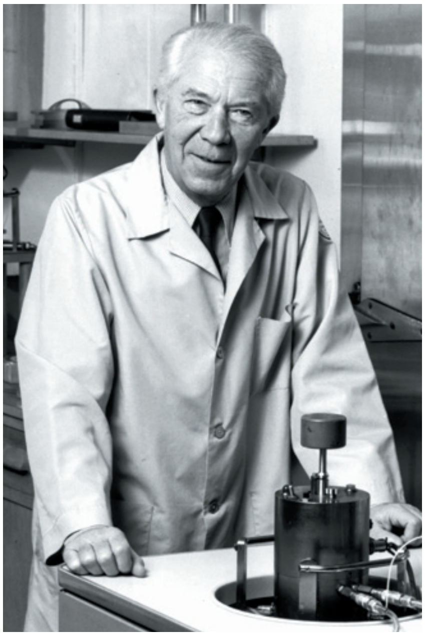
The path to excellence in experimental biology is long and arduous. Christian de Duve was one of the few to reach a summit that opened new vistas onto uncharted territories. In 1974, he shared the Nobel Prize in Physiology or Medicine with Albert Claude and George Palade, for their discoveries "concerning the structural and functional organization of the cell".

De Duve was born in 1917 near London, where his parents, of Belgian–German ancestry, had moved to escape the First World War. Soon after the war ended, the family returned to Antwerp in Belgium, where Christian attended school — at which he excelled. He then enrolled as a medical student at the Catholic University of Louvain in 1934 and joined its physiology laboratory, directed by J. P. Bouckaert. In 1943, he married Janine Herman, who subsequently became a noted painter. She steadfastly supported him throughout the rest of her life. His many moves required huge sacrifices from her and their four children, and gave de Duve the freedom to focus his efforts on scientific research.

During these years, de Duve made several respectable discoveries about storage and retrieval of the body's principal fuel, glucose, as affected by the pancreatic hormones insulin and glucagon. After the Second World War, top-notch international laboratories opened their doors to him. He first went to Hugo Theorell's lab at the Nobel Medical Institute in Stockholm and then to Gerty and Carl Cori's lab at Washington University in St. Louis.

In 1947, family ties enticed de Duve to return to Louvain as a professor in its medical school. On his way back from St Louis, he stopped at the Rockefeller Institute for Medical Research (now Rockefeller University) in New York for a life-changing meeting with his countryman Albert Claude.

Claude had gone to Rockefeller in 1929 to isolate what we now know as the Rous sarcoma virus. On route to this goal, he revolutionized cell research by experimenting with two recently developed instruments: the electron microscope and the high-speed centrifuge. The electron microscope enabled Claude, with Keith Porter, to look inside cultured cells at a magnification much greater than that possible with a light microscope. Their epochal discovery, reported in 1945, was the lace-like network surrounding the cell's nucleus, the 'endoplasmic reticulum'. Using the high-speed centrifuge, Claude and his collaborators managed to separate and enrich various cell components on the basis of their size and density.



INGBERT GRUTTNER/THE ROCKEFELLER UNIV.

Heading his own laboratory in Louvain, de Duve continued research on insulin and glucagon, this time inspired by Claude. Among his first observations at Louvain was that the enzyme activity of glucose-6-phosphatase was mostly associated with a cell fraction that sedimented at high centrifugal forces. More detailed analyses by de Duve, and by Philip Siekevitz and Palade, established glucose-6-phosphatase as the first 'marker' enzyme for the endoplasmic reticulum. These findings supported the idea that the cell contains distinct compartments with characteristic enzyme activities.

De Duve promptly set his insulin research aside to search for these uncharacterized cellular compartments. Because it could be readily measured, he chose the enzyme acid phosphatase. His observations — which to most contemporaries would have seemed trivial curiosities related to this enzyme's activity and its cellular partitioning, and thus not worth close attention — became important clues for de Duve. Could acid phosphatase be a marker enzyme for yet another cellular compartment? Perhaps the enzyme was part of a cell's digestive quarter, stored with other hydrolytic enzymes to break down the multitude of macromolecules taken up by the cell?

To answer these questions, he used an early version of an '-omic' approach. He combed the literature for other known hydrolytic enzymes and, after several reassuring findings, went public with the concept of the lysosome (C. de Duve *et al. Biochem. J.* **60,** 604–617; 1955), now known as the hub of the cell's digestive system. Using similar approaches, his laboratory also discovered the peroxisome, a cellular compartment containing enzymes involved in oxidation.

Thereafter, others discovered lysosomal and peroxisomal diseases. Some of these disorders, in which one lysosomal enzyme is either missing or faulty, can now be treated by providing the requisite lysosomal enzyme.

Next, de Duve expanded and consolidated these findings and established himself as a skilled administrator. He directed a lab at Rockefeller University (1962–87) and founded and ran (1975–85) the International Institute of Cellular and Molecular Pathology in Brussels (a prescient realization of the idea of translational medicine), subsequently named the de Duve Institute. He also had a principal role in creating the prestigious and unique L'Oreal–UNESCO Awards for Women in Science, presented to one woman annually in each of five continents. In the last 'contemplative' period of his life, he wrote influential books directed at the educated lay public, primarily on biology and evolution.

As de Duve wrote in his memoirs, *Sept vies en une* (Odile Jacob Sciences, 2013), he was a precocious child, perpetually the best student ('*primus perpetuus*') in his school except for one year, when he was declared out of competition ('*hors concours*') so that another student could come top. These early accolades, and the many that followed, reinforced an inherent sense of self-confidence that sometimes had unintended consequences. He emphatically denied having been an authoritarian boss ('*patron autoritaire*'), but in a typically de Duvean way admitted, in brackets, that some people may feel differently. Personally, he struck me as a warm, humorous and compassionate human being, whom I will miss thoroughly. ■

Günter Blobel *is professor of cell biology and a Howard Hughes Medical Institute investigator at the Rockefeller University, New York, USA, where he knew de Duve.*
e-mail: blobel@mail.rockefeller.edu

FORUM: Atmospheric science

The seeds of ice in clouds

An investigation of droplet freezing in clouds suggests that a minor component of mineral dust in the atmosphere is the main catalyst for this process. Two experts discuss the ramifications of this finding for those investigating cloud–droplet freezing, and for scientists studying atmospheric aerosols. SEE LETTER P.355

THE PAPER IN BRIEF

- Mixed-phase clouds contain both liquid water droplets that have been 'supercooled' to temperatures below the freezing point of water and ice particles.
- The amount of ice affects many of these clouds' properties, such as their extent and lifetime.
- Aerosol particles of minerals in the atmosphere, known as ice nuclei, catalyse the freezing of cloud droplets.
- Atkinson *et al.*[1], as they report on page 355 of this issue, have studied ice nucleation in conditions found in mixed-phase clouds*.
- They find that ice nucleation is dominated by feldspar minerals rather than clay minerals, as had been generally thought.

Rare but active

THOMAS KOOP

Earth's weather and climate are governed by clouds whose properties are largely determined by aerosol particles and by physical processes that occur at the micrometre to metre scale[2]. Although the basics of cloud formation are well understood, quantitatively predicting cloud properties remains challenging, partly because of the chemical diversity and varying abundance of aerosol particles. Atkinson and co-workers have helped to address this problem by considering the effects of different kinds of mineral particles on ice nucleation in clouds.

In many clouds, tiny water droplets supercool significantly before heterogeneous ice nucleation is triggered by ice nuclei, a process known as ice activation. The resulting frozen cloud droplets grow into larger ice particles, which may then initiate precipitation (Fig. 1). This cascade of processes depends crucially on local concentrations of ice nuclei and on the temperature at which these nuclei activate ice.

There are two approaches for establishing a global representation of ice-nuclei concentrations and ice-activation temperatures. The first is to use *in situ* measurements to construct an empirical relationship of how the concentration of active ice nuclei varies with temperature[3]. The second approach, and that adopted by Atkinson *et al.*, is to perform laboratory studies with aerosol particles for which the atmospheric abundance is known[4,5]. The authors froze micrometre- and millimetre-sized water droplets containing ice nuclei at temperatures typical of mixed-phase clouds (about 250–265 kelvin). Surprisingly, they found that potassium-rich feldspar seems to be the most active mineral for ice nucleation, even when its low abundance of just a few per cent by mass in soil-mineral dusts — the main source of mineral dust in the atmosphere — is taken into account.

However, when the authors combined a global model of atmospheric feldspar-dust concentration with their laboratory results, the predicted concentrations of feldspar ice nuclei did not fully agree with those derived from field studies, indicating the need for improved dust modelling, or for alternative ice nuclei to be identified in certain regions. Moreover, the authors argue that ice-nucleation activities reported in earlier studies of more-abundant clay-mineral dusts could have been elevated by traces of feldspar — a proposition that will spark controversy, as well as more attentive characterization of ice nuclei in future experiments.

The results of field and laboratory studies should eventually converge towards a consistent representation of atmospheric ice nuclei that is sufficient for cloud modelling. But for now, it is disappointing to acknowledge how little fundamental understanding we have about heterogeneous ice nucleation. For example, what makes a good ice nucleus? Initial proposals suggested that the lattice of an ice nucleus must match that of ice, as in the 'classical' case of silver iodide. But such lattice matching does not seem to be a reliable predictor of ice-nucleation activity[6]. Moreover, a wide variety of non-crystalline ice nuclei have been identified, for example amorphous solids, pollen and bacteria, surfactant monolayers, and even dissolved polymer molecules and proteins.

Resolving the unknowns will surely require the efforts of collaborative multidisciplinary consortia to combine the results obtained from elaborate experimental tools with those of theoretical simulations. Even then, it could be decades before a clear picture of heterogeneous ice nucleation emerges. The results will not only benefit the atmospheric sciences, but will also have applications in engineering, for example the cryostorage of biological tissues and food, and in the development of methods for preventing aircraft icing up or pipes freezing. In the meantime, further surprising discoveries such as that of Atkinson *et al.* are likely to be made.

Thomas Koop *is in the Faculty of Chemistry, Bielefeld University, Bielefeld 33615, Germany.*
e-mail: thomas.koop@uni-bielefeld.de

Not all dust is equal

NATALIE MAHOWALD

Atkinson *et al.*[1] suggest not only that the effects of mineral aerosols dominate those of other ice nuclei, as has been reported elsewhere[7], but also that feldspar is the most effective ice nucleus of all. The importance of aerosol–cloud interactions to climate[8] means that the authors' work is a call to arms for scientists who study atmospheric aerosols, and in particular for those who study dust.

Atmospheric aerosols are highly heterogeneous in space and time, and so measurements and models of these aerosols tend to simplify their complexity. Researchers feel lucky if they can obtain sufficiently long time series at enough locations to characterize the variability; even fewer measurements include

*This article and the paper under discussion[1] were published online on 12 June 2013.

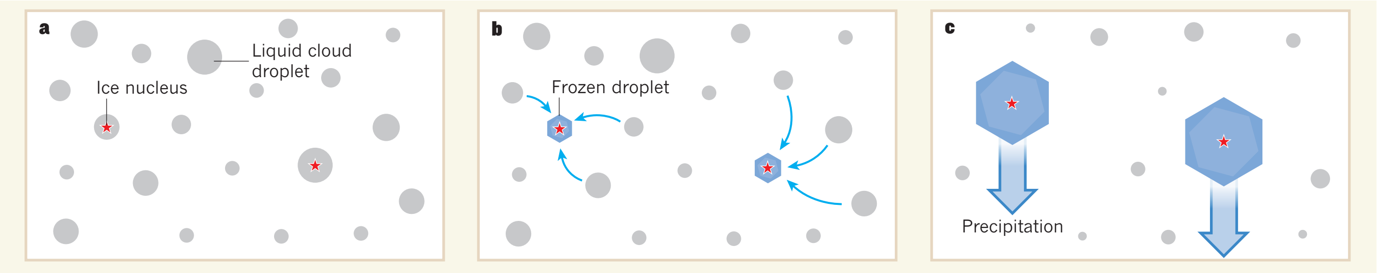


Figure 1 | Ice formation and precipitation. **a**, **b**, Aerosol particles known as ice nuclei catalyse the formation of a few frozen cloud droplets (typically several micrometres in diameter) from supercooled droplets of liquid water. These ice crystals grow at the expense of the remaining majority of liquid droplets, through transfer of water vapour (blue arrows). **c**, The resulting large ice particles (often several tens to more than 100 micrometres in diameter) have higher fall velocities than the small liquid droplets, and may initiate precipitation. Atkinson *et al.*[1] report that feldspar particles are the most effective mineral ice nuclei.

details of aerosol composition. Rarer still are studies that, as well as estimating the fraction of aerosols that is composed of minerals, also analyse what the different minerals are. Similarly, because of the expense of simulating the effects of many types of minerals and the lack of comprehensive data, the incredible variability of mineral aerosol composition is ignored in climate models. Instead, mineral aerosols are usually modelled together, as a bulk dust. Any atmospheric processing of mineral aerosols that would modify their chemical and physical properties is also commonly ignored in models.

Atkinson and co-workers' findings demonstrate the need for more observations of the mineralogical composition of mineral aerosols; currently, such observations are few and far between (see the Supplementary Information of the paper[1]). More information about the effects of acids on mineral aerosols is also required to gauge the role of these reactions in the atmospheric processing of minerals. For example, do acids convert feldspar into less-effective ice nuclei, such as clays? We also need a better understanding of the distribution of minerals in areas of soil that act as sources of dust. In addition, we must learn more about how humans and climate have changed, and will change, desert dust (and feldspar dust in particular) over time. The limited evidence available suggests that the mass of dust worldwide doubled over the twentieth century[9].

Finally, Atkinson and colleagues' work requires us to rethink how aerosols and aerosol–cloud interactions are modelled: multiple types of minerals, as well as their chemical reactions with compounds such as sulphates or organic acids in the atmosphere, must be considered. This means that substantial increases in the complexity and computational expense of models are needed. Scientists should consider whether we can use a proxy for the potential of different mineral compositions to nucleate ice — instead of the effects of specific minerals — to reduce the complexity of the problem such that mineral aerosols can be included in computationally expensive climate models more correctly.

In retrospect, the finding that a specific mineral is responsible for most ice-nucleation events in mixed-phase clouds is perhaps not that surprising, because the chemical and physical properties of different mineral aerosols are so disparate. For instance, earlier studies have highlighted the importance of aerosol mineralogy in the interactions of atmospheric dust with light[10] and in ocean biogeochemistry[11]. Nevertheless, Atkinson and colleagues' discovery is extremely important: when it comes to ice nucleation, not all dust is created equal. ■

Natalie Mahowald *is in the Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York 14850, USA.*
e-mail: nmm63@cornell.edu

1. Atkinson, J. D. *et al.* **498,** 355–358 (2013).
2. Baker, M. B. & Peter, T. *Nature* **451,** 299–300 (2008).
3. DeMott, P. J. *et al. Proc. Natl Acad. Sci. USA* **107,** 11217–11222 (2010).
4. Niemand, M. *et al. J. Atmos. Sci.* **69,** 3077–3092 (2012).
5. Hoose, C. & Möhler, O. *Atmos. Chem. Phys.* **12,** 9817–9854 (2012).
6. Croteau, T., Bertram, A. K. & Patey, G. N. *J. Phys. Chem. A* **112,** 10708–10712 (2008).
7. Cziczo, D. *et al. Science* http://dx.doi.org/10.1126/science.1234145 (2013).
8. Forster, P. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 130–234 (Cambridge Univ. Press, 2007).
9. Mahowald, N. *et al. Atmos. Chem. Phys.* **10,** 10875–10893 (2010).
10. Sokolik, I. N. & Toon, O. B. *J. Geophys. Res.* **104,** 9423–9444 (1999).
11. Journet, E. *et al. Geophys. Res. Lett.* **35,** L07805 (2008).

BEHAVIOURAL BIOLOGY

Archaeology meets primate technology

A study of wild capuchin monkeys that crack nuts using stone hammers reveals temporal and spatial patterning of the relics of their technological efforts, confirming that such behaviours can be studied from an archaeological perspective.

ANDREW WHITEN

Our ancestors have been fashioning and using stone tools for at least 2.5 million years[1]. Bronze blades began to replace lithic axes a mere few thousand years ago, so percussive stone tools — hammers and axes that function through targeted force — have characterized more than 99.9% of human technological evolution[2]. The discovery of stone-tool use in other primates has offered exciting opportunities to examine such behaviour in living species. West African chimpanzees have provided the focus for this research for 30 years, but it was revealed a decade ago[3] that bearded capuchin monkeys (*Sapajus libidinosus*) also use hammer stones to crack nuts (Fig. 1a). Writing in the *Journal of Archaeological Science*, Visalberghi *et al.*[4] present the fruits of an interdisciplinary project[5] that emerged from this discovery. Their study goes beyond behavioural observations to log the archaeological signatures of percussive tool use by capuchins.

This research represents one of the first comprehensive empirical examples of the new discipline of primate archaeology[6–8]. Working in the monkeys' open, savannah-like woodland habitat in Brazil, the authors located 58 active nut-cracking anvil sites, which they identified by the presence of hammer stones, pits on the stone or wooden-log anvils, and nut remains. Each month for three years, the scene at each anvil was inspected and photographed,

A, ELISABETTA VISALBERGHI; B, MICHAEL D. GUMERT; C, PRIMATE RESEARCH INST. OF KYOTO UNIV.; D, DIETRICH STOUT

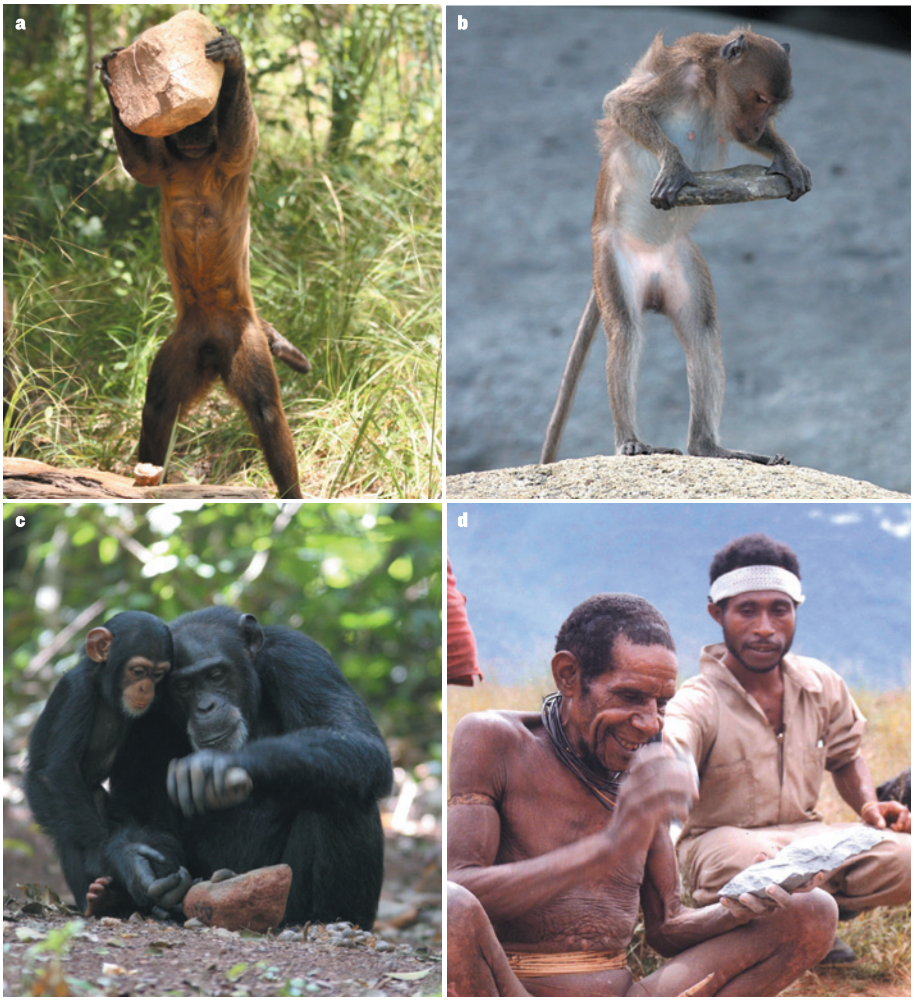


Figure 1 | Use of percussive stone tools in primates. **a**, A male capuchin monkey weighing just over 4 kilograms uses a 3.5-kg stone to crack a highly resistant piassava nut. Hammer stones typically weigh around 1 kg. **b**, A long-tailed macaque raises a stone to crack an intertidal snail. **c**, A common chimpanzee cracking nuts using a stone hammer and anvil. **d**, People knapping stone tools on Irian Jaya, Indonesia.

the nuts were cleared, the hammers replaced and the array was re-photographed to track material changes at the site.

The authors found a median usage per anvil of 35% of months, and a maximal use of a single anvil in 30 out of 36 months. Hammer transport was a relatively rare occurrence, with just 40 cases of hammers being shifted farther than 3 metres in 1,872 visits. However, on seven of these occasions, the hammer was moved up to 10 m away to a boulder that had not previously been used as an anvil. Moreover, in four cases, viable new hammer stones, which are quite rare at the site, appeared at the inspected anvils. And a hammer disappeared from the site on 17 occasions, in two cases being returned 1 and 5 months later.

Putting these and previous observations[9] together, it seems likely that rare but more extended transport between anvils may also occur; the researchers are planning longer-term recordings, and it will be interesting to see what these reveal. Alongside other elements in this study, such as records of the weathering of nut-case remains, the observations begin to delineate the material effects of a non-human primate's technological activities on the landscape in both space and time, as well as indirectly charting large-scale patterns in the monkeys' tool-related behaviour.

Parallel studies on chimpanzees are under way[10]. Do studies such as these merit the 'archaeology' epithet their authors promote? Dictionary definitions suggest not, referring instead to studies of "man's past" and "ancient cultures". Indeed, we tend to think of archaeologists as digging deep to find crucial remains. It is true that the remains examined by Visalberghi and colleagues are far from ancient, although, in the case of chimpanzee nut-cracking, evidence of a history back to 4,300 years has been excavated[11]. However, such definitional quibbles can be seen as pedantic. Extending the scope of human-focused disciplines to other species has yielded insight in several domains of evolutionarily focused enquiry, culture itself being one of them[2].

A key question that must be addressed by such studies is thus whether capuchin technology is culturally transmitted, through observational learning. Controlled experiments have demonstrated that alternative foraging techniques that are seeded in different captive groups of capuchins spread through social learning to become traditions[12]. Such experiments are hard to emulate in the wild, but Visalberghi and co-workers' findings offer a variety of circumstantial evidence for cultural transmission, which the authors believe is supported by the correlated presence at anvils of arrays of key materials[13]. These findings are complemented by field experiments[14] that elegantly demonstrate a sophisticated understanding of optimal tool properties, such as mass and size, in capuchin monkeys.

Humans and capuchins are separated from their common ancestor by about 35 million years. So, can studying these monkeys influence our understanding of the lithic technology that pervaded so much of our own evolutionary history[2,15]? I believe so. We know that the long-tailed macaque also uses stone hammers to process hard-shelled foods, such as oysters and sea snails, on rocky shorelines (Fig. 1b)[16]. The shells acquire different wear patterns as a result of the monkeys' use of different tools for these various targets. Macaques are Old World monkeys, the group of primate species that are today found in Africa and Asia, as opposed to the capuchins, which belong to the New World monkeys of Central and South America. Together, these studies suggest that using stone hammers to access embedded foods may be a widespread but often latent capability among monkeys as well as apes, which finds expression in response to the co-occurrence of a small set of facilitating circumstances. The convergence on these behaviours by such diverse species of primate offers opportunities to identify ecological and other factors that support the emergence of percussive stone technology. For example, intriguing findings are already emerging in the capuchin studies that contradict a popular hypothesis that percussive tool use functions to overcome seasonal food scarcity[17].

The form that percussive, lithic technology takes in the chimpanzee — the species with whom we shared our most recent common ancestor — may have further significance. Whereas a capuchin generally needs to rear bipedally to use a stone to crack nuts (Fig. 1a), chimpanzees typically sit, truncally erect, and may use one hand to wield the hammer and the other to manipulate the target (Fig. 1c). In a common ancestor, this configuration would have provided a preadaptation to the approach used by modern human stone knappers[18] (Fig. 1d). ■

Andrew Whiten *is in the Centre for Social Learning and Cognitive Evolution, School of Psychology and Neuroscience, University of St Andrews, St Andrews KY16 9JP, UK.*
e-mail: aw2@st-andrews.ac.uk

1. McPherron, S. P. *et al. Nature* **466,** 857–860 (2010).
2. Whiten, A., Hinde, R. A., Laland, K. N. & Stringer, C. B. *Phil. Trans. R. Soc. B* **366,** 938–948 (2011).
3. Fragaszy, D., Izar, P., Visalberghi, E., Ottoni, E. B. & de Oliviera, M. G. *Am. J. Primatol.* **64,** 359–366 (2004).
4. Visalberghi, E., Haslam, M., Spagnoletti, N. & Fragaszy, D. *J. Archaeol. Sci.* **40,** 3222–3232 (2013).
5. www.ethocebus.net
6. McGrew, W. C. & Foley, R. A. *J. Hum. Evol.* **57,** 335–336 (2009).
7. Haslam, M. *et al. Nature* **460,** 339–344 (2009).
8. Wynn, T., Hernandez-Aguilar, R. A., Marchant, L. F. & McGrew, W. C. *Evol. Anthropol.* **20,** 181–197 (2011).
9. Visalberghi, E. *et al. Primates* **50,** 95–104 (2009).
10. Carvalho, S., Biro, D., McGrew, W. C. & Matsuzawa, T. *Anim. Cogn.* **12** (Suppl.), 103–114 (2009).
11. Mercader, J. *et al. Proc. Natl Acad. Sci. USA* **104,** 3043–3048 (2007).
12. Dindo, M., Thierry, B. & Whiten, A. *Proc. R. Soc. Lond. B* **275,** 187–193 (2008).
13. Fragaszy, D. *et al. Phil. Trans R. Soc. B* (in the press).
14. Visalberghi, E. *et al. Curr. Biol.* **19,** 213–217 (2009).
15. Goren-Inbar, N., Sharon, G., Melamed, Y. & Kislev, M. *Proc. Natl Acad. Sci. USA* **99,** 2455–2460 (2002).
16. Gumert, M. D., Kluck, M. & Malaivijitnond, S. *Am. J. Primatol.* **71,** 594–608 (2009).
17. Spagnoletti, N. *et al. Anim. Behav.* **83,** 1285–1294 (2012).
18. Whiten, A., Schick, K. & Toth, N. *J. Hum. Evol.* **57,** 420–435 (2009).

HIV

Integration triggers death

That HIV cripples the immune system by killing CD4$^+$ T cells has long been known. It now emerges that the protein DNA-PK, activated by viral integration into the host-cell genome, is the agent of this death response. SEE LETTER P.376

ANNA MARIE SKALKA

Retroviruses, the class of virus that includes HIV, do not ordinarily destroy the cells that they infect. Instead, they are propagated essentially as genetic parasites: after a DNA copy of the retrovirus's own RNA genome has been integrated into the DNA of a host cell, the cell is exploited to express viral molecules, and progeny viruses are released even as the host cell continues to thrive. But infection of human white blood cells known as activated CD4$^+$ T cells is a marked exception. In fact, it is the en masse killing of these cells by HIV that gives rise to the severe immunodeficiency that is AIDS. In this issue, Cooper *et al.*[1] (page 376) report that this death is the T cells' response to the attack on its genome by the viral integration machinery. Furthermore, the authors reveal that the main player in this response is DNA-dependent protein kinase, an enzyme normally associated not with cell death, but with the repair of DNA damage*.

Retroviruses enter their host cells fully equipped to carry out the first essential steps of viral replication. The viral enzyme reverse transcriptase ensures rapid synthesis of a double-stranded DNA copy of the viral RNA genome. Another virus-associated enzyme, integrase, then binds to and processes the ends of this viral DNA molecule as soon as they are formed. A protein complex containing the viral DNA, integrase and other viral and host proteins is subsequently transported to the cell's nucleus, where integrase catalyses a concerted cleavage and joining reaction in which the 3′ ends of the viral DNA are joined to the 5′ ends of a double-stranded cut in the host DNA. The remaining gaps and overhangs at this integration site are probably repaired by the cell within 26 hours of infection, when expression of viral capsid proteins can be detected[2].

Cooper and colleagues studied this process in human CD4$^+$ T cells infected with HIV *in vitro*. They observed that the cells express viral DNA and HIV-encoded proteins within 36 hours, but that this expression ceases by the second day, concomitant with massive cell death. This time frame is consistent with the estimated half-life of infected activated CD4$^+$ T cells in patients with AIDS[3], but the trigger for the death of these cells had not previously been identified. Cooper *et al.* found that both death and loss of HIV-protein expression in these cells could be prevented by the addition of inhibitors of viral reverse transcriptase or integrase (the authors used efavirenz and raltegravir, respectively) before infection. Furthermore, raltegravir treatment of activated CD4$^+$ T cells isolated from infected individuals (before therapy with antiretroviral drugs) rescued some of these cells from virus-induced death.

The authors also demonstrate that an increase in the prevalence of free viral DNA ends, which occurs when integration is blocked, does not, as has been suggested previously[4], promote cell killing. And they found that the expression of viral genes following integration into the host genome also does not promote cell death. These and other results from their study strongly support the conclusion that it is the viral-integration step that promotes killing of HIV-infected T cells.

This proposal is consistent with the notion that integration of virus-derived DNA is perceived by the cell as a DNA-damaging event[5]. Mammalian cells have evolved intricate and partially overlapping mechanisms for responding to DNA damage, and three members of the phosphatidylinositol-3-kinase-like protein family — ATM, ATR and DNA-dependent protein kinase (DNA-PK) — are central to this response[6]. Both ATM and DNA-PK are known to be recruited to and activated by double-stranded DNA breaks, and previous studies have implicated these enzymes in

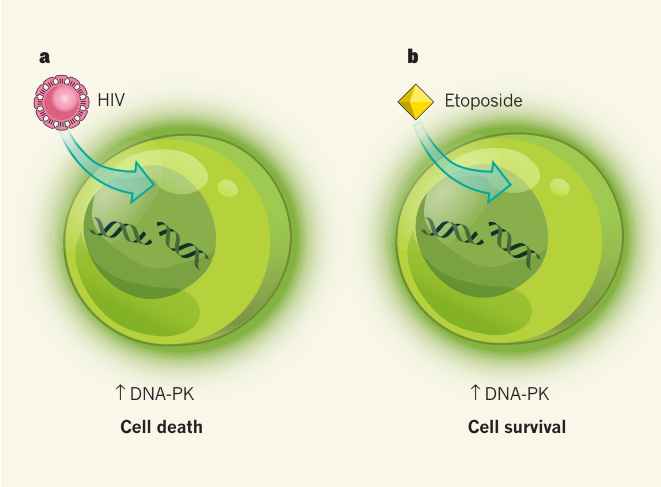


Figure 1 | Opposing responses to DNA damage. a, When a cell is infected with HIV, the virus's integrase enzyme induces double-stranded breaks in the cell's DNA and inserts a DNA copy of its RNA genome. Some cells survive this process, but activated CD4$^+$ T cells do not. Cooper *et al.*[1] reveal that this is because, in these cells, subsequent activation of DNA-PK leads to p53-mediated apoptotic cell death. **b,** Treatment of activated CD4$^+$ T cells with the agent etoposide also leads to double-stranded DNA breaks, by interacting with the DNA-unwinding protein topoisomerase II and preventing the rejoining of DNA breaks that occur during normal cell replication. In this case, however, DNA-PK activation promotes DNA repair, allowing the cells to survive.

*This article and the paper under discussion[1] were published online on 5 June 2013.

DNA repair following the integration of viral DNA from HIV and other retroviruses in several cell types[5,7,8]. It may therefore not have been surprising for Cooper and colleagues to discover that HIV infection of $CD4^+$ T cells stimulated both ATM and DNA-PK activity — but the role they uncovered for DNA-PK was a surprise.

The researchers also observed increased levels of two other indicators of the cellular response to DNA damage: phosphorylation of the histone protein H2AX and of the protein p53. There is increasing evidence that, in response to severe DNA damage or abnormal DNA structures, DNA-PK may form a complex with p53 that promotes cell death by apoptosis[9]. Cooper *et al.* provide evidence that this is indeed the case in HIV-infected activated $CD4^+$ T cells, by showing that treatment of these cells with inhibitors of p53 or DNA-PKcs (the catalytic subunit of DNA-PK), but not with an ATM inhibitor, blocked cell killing in response to viral infection.

This result is made all the more striking by the authors' report that the use of the same DNA-PKcs inhibitors increased the death of activated $CD4^+$ T cells that were not infected but instead had been treated with etoposide, an agent that induces single- and double-stranded DNA breaks by inhibiting the strand-joining activity of the protein topoisomerase II. Thus, it seems that DNA-PK activity promotes opposite effects in these cells depending on the context: in response to insults inflicted by etoposide, DNA-PK is required for DNA repair and cell survival, whereas following integration of HIV DNA, it induces apoptosis (Fig. 1).

This study convincingly fingers HIV integrase and DNA-PK as crucial players in the death of $CD4^+$ T cells following HIV infection, but it leaves us hungry for mechanistic insight. How can the seemingly conflicting roles of DNA-PKcs in these cells be reconciled?

One clue might be that the DNA lesions introduced by etoposide and HIV integration are quite different, with the latter involving intermediate single-stranded attachments of viral DNA. However, because HIV proteins and even copious numbers of progeny viruses are produced in the first day or two after infection of $CD4^+$ T cells, we may assume that the integration-induced damage is actually repaired in these cells: apoptosis seems to be a delayed response.

Another clue may come from the cell-type specificity of this response — HIV is not toxic to other human cells, such as macrophages, or to numerous cell lines. Perhaps the concentration or cellular location of DNA-PK or its targets in different cell types affects the outcome of its activation. Alternatively, HIV-infected T cells might experience an imbalance in other components of the repair machinery[10] that tips the cells towards apoptosis. But despite these remaining questions, Cooper and colleagues' suggestion that treatment with drugs that block HIV integration will not only inhibit virus replication, but should also enhance T-cell survival, seems spot on. ■

Anna Marie Skalka *is at the Fox Chase Cancer Centre, Philadelphia, Pennsylvania 19111, USA.*
e-mail: am_skalka@fccc.edu

1. Cooper, A. *et al. Nature* **498,** 376–379 (2013).
2. Vandegraaff, N., Kumar, R., Burrell, C. J. & Li, P. *J. Virol.* **75,** 11253–11260 (2001).
3. Ho, D. D. *J. Clin. Invest.* **99,** 2565–2567 (1997).
4. Li, L. *et al. EMBO J.* **20,** 3272–3281 (2001).
5. Skalka, A. M. & Katz, R. A. *Cell Death Differ.* **12,** 971–978 (2005).
6. Yang, J., Yu, Y., Hamrick, H. E. & Duerksen-Hughes, P. J. *Carcinogenesis* **24,** 1571–1580 (2003).
7. Daniel, R. *et al. Mol. Cell. Biol.* **21,** 1164–1172 (2001).
8. Lau, A., Kanaar, R., Jackson, S. P. & O'Connor, M. J. *EMBO J.* **23,** 3421–3429 (2004).
9. Hill, R. & Lee, P. W. K. *Cell Cycle* **9,** 3460–3469 (2010).
10. Shao, L., Goronzy, J. J. & Weyand, C. M. *EMBO Mol. Med.* **2,** 415–427 (2010).

BIOMIMETICS

Flying like a fly

When biologists unravelled the principles of insect flight, they inspired a generation of engineers to build on their aerodynamic feats. Thanks to a revolution in micro-manufacturing techniques, the first robotic fly now flies.

DAVID LENTINK

A robotic fly discreetly monitoring our homes and command centres was a pop-cultural manifestation of cold-war paranoia. It was also pure fiction, because scientists of the time were unable to explain the mechanics of insect flight. Relying on aerodynamic theory that was appropriate for fixed-wing aircraft, their calculations could infer only that insect wings generate too little lift to remain aloft. But our understanding of insect aerodynamics, and ability to build robots that mimic and exploit it, has increased immensely over the past two decades. A culmination of this is the report by Wood and colleagues[1] (Ma *et al.*) in *Science* of the first controlled flight of an at-scale robotic fly*.

An important step on the way to elucidating the secrets of insect flight came in 1996 from researchers using a gigantic, dynamically scaled model hawkmoth[2]. This robot flapped its wings at a stately frequency of only once every 3 seconds, which was calculated to reproduce the airflow and lift force of a hovering moth. By releasing smoke from within the wings, the authors were able to visualize a tornado-like vortex that ran outwards along the leading edge of each wing. The remarkable stability of this leading-edge vortex enables the wings of insects to operate at angles of attack at which the wings of an aircraft stall, and consequently to generate more lift.

Although the smoke experiments revealed the vortex, they did not quantify the extra lift. To measure this force, another group[3] created the Robofly, a robotic fruitfly enlarged by a factor of 100, which they studied submerged in a tank of mineral oil. Fluid force is proportional to the ratio of viscosity-squared to density, and for otherwise-similar flow patterns this force is 50,000 times greater in oil than in air. This

*This News & Views article was published online on 12 June 2013.

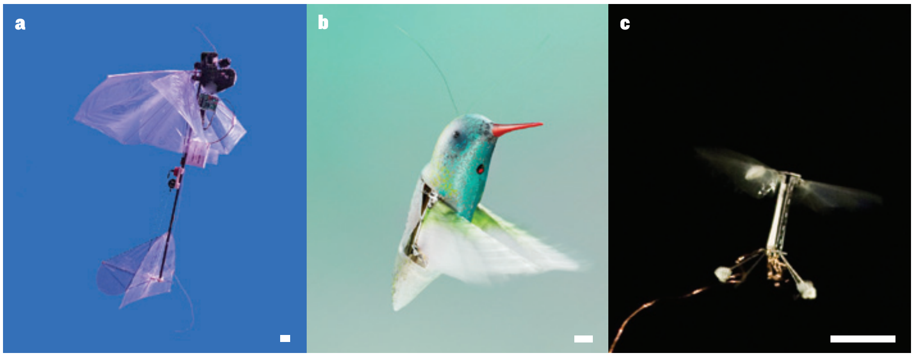


Figure 1 | Winged victories. Three successive iterations of miniaturized robots that each mimic certain aspects of animal hovering flight. **a**, The passively stable DelFly[7] hovers like an insect that is controlled by its tail. **b**, The tailless Nano Hummingbird[8] is stabilized by an on-board autopilot, which controls the wings' angle in a way analogous to that seen in real hummingbirds. **c**, Ma and colleagues' robot fly[1], shown here on its maiden flight, is controlled by a tether that provides modulated power to each flight 'muscle' of the wing. Scale bars, 10 millimetres (estimated).

amplification allowed the authors to record and disentangle the myriad aerodynamic mechanisms that fruitflies exploit to perform their intricate hovering flight manoeuvres.

Insight from the Robofly enabled electrical engineers to design at-scale robotic flies[4] that researchers had previously only imagined[5], kicking off robot-fly evolution. At the time, fly-weight robots were an engineer's fabrication nightmare, because electronic components were heavy. No wonder, then, that the first flapping robot that hovered like a fly — the 360-millimetre-wingspan Mentor — weighed more than 400 grams[6]. This weight limited Mentor to short vertical and hovering flights, which were stabilized by an autopilot. The next incarnation, the 280-mm DelFly (Fig. 1a), which relied on passive stability instead of bulky electronics, weighed only 16 g and could fly for 16 minutes[7]. DelFly performed vertical take-offs and landings, hovered, and flew forward like a dragonfly.

Subsequently, a young hobbyist scaled this design down to a mere 60-mm wingspan and 930-mg mass, and flew it indoors. A 2009 video posted on YouTube (go.nature.com/qhrbnl) demonstrates the remarkable battery-powered flight, lasting more than 1 minute, of this robot, which was developed at a time when competing multimillion-dollar research projects that aimed to achieve similar results could not get off the ground. But things changed with the Nano Hummingbird[8], the first tailless flapping robot that could take off and land vertically (Fig. 1b). Measuring 160 mm, the robot can fly for 11 minutes on battery power, is stabilized by an autopilot and steers by controlling the angle of attack over the course of each wingbeat — just like real hummingbirds, which have been dubbed nature's honorary insects. The robot's extreme manoeuvrability is comparable to that of hummingbirds and flies. On the flip side, it still weighs 5 times more than common species of hummingbird and 1,000 times more than a house fly.

These robots confirmed the experimental prediction that flapping flyers could be scaled down to insect size and still function; fundamentally, this is because the aerodynamic mechanisms that underlie their flight are not limited by scale[7]. However, further miniaturization was prevented by the absence of efficient lightweight fabrication technology at the millimetre scale. But researchers in the Wood laboratory have spent more than a decade devising ways to bridge this technological gap. The group last year reported a revolutionary millimetre-scale manufacturing technique, inspired by pop-up books, that can mass-produce 30-mm fly-like robots weighing only 80 mg[9]. To get around the implacable scaling laws that degrade the performance of electric motors and bearings at this scale, the team also developed efficient replacements in the form of miniature piezoelectric actuators and low-friction flexible joints.

These advances led to the remarkable realization of Wood and colleagues' at-scale robot fly (Fig. 1c). However, the device comes with strings attached: a tether connects the robot to a grounded battery and autopilot. The latter monitors and adjusts the flight path of the robot almost beat by beat. Although micrometre-scale on-board autopilot is close to completion, the development of microbatteries remains remarkably challenging. Radically new battery technology is needed to power this wave of free-flying, flapping microrobots out of science fiction and into contemporary society.

When this occurs, insect-sized robots will probably be used first as inconspicuous (and inexpensive) eyes in the skies to help us to obtain situation awareness, for example during hostage situations or in urban war zones, and later perhaps as artificial agricultural pollinators. Ma *et al.* suggest that their robot fly will also advance our biological understanding of insect flight. The robot could, for example, be manipulated to test specific hypotheses that concern stability and control. Unfortunately, the flapping wings of the robot will not push the boundaries of aerodynamic efficiency — in one-on-one comparisons, helicopter rotors consistently require less power, based on weight, than flapping wings[7,8]. Flapping robots are, however, poised to fly more robustly in cluttered and turbulent environments. Here, whereas animals succeed, the current generation of microdrones fails drastically. Perhaps soldiers of the future will need to carry a swatter on the battlefield. ■

David Lentink *is in the Department of Mechanical Engineering, Stanford University, Stanford, California 94305, USA.*
e-mail: dlentink@stanford.edu

1. Ma, K. Y., Chirarattananon, P., Fuller, S. B. & Wood, R. J. *Science* **340,** 603–607 (2013).
2. Ellington, C. P., van den Berg, C., Willmott, A. P. & Thomas, A. L. R. *Nature* **384,** 626–630 (1996).
3. Dickinson, M. H., Lehmann, F. O. & Sane, S. P. *Science* **284,** 1954–1960 (1999).
4. Fearing, R. S. *et al.* in *Proc. IEEE Int. Conf. Robotics Automation* 1509–1516 (2000).
5. Flynn, A. M. in *Proc. IEEE Micro Robots and Teleoperators Workshop* 221–225 (1987).
6. Zdunich, P. *et al. J. Aircraft* **44,** 1701–1711 (2007).
7. Lentink, D., Jongerius, S. R. & Bradshaw, N. L. in *Flying Insects and Robots* (eds D. Floreano *et al.*) 185–205 (Springer, 2010).
8. Keennon, M., Klingebiel, K., Won, H. & Andriukov, A. in *50th AIAA Aerospace Sciences Meeting* **0588,** 1–24 (2012).
9. Sreetharan, P., Whitney, J., Strauss, M. & Wood, R. *J. Micromech. Microeng.* **22,** 055027 (2012).

PHYSICS

Heavy calcium nuclei weigh in

The configurations of calcium nuclei make them good test cases for studies of nuclear properties. The measurement of the masses of two heavy calcium nuclei provides benchmarks for models of atomic nuclei. SEE LETTER P.346

ALEXANDRA GADE

Perhaps the most fundamental observable property in nuclear physics is the mass of an atomic nucleus, which is related to how strongly the protons and neutrons within are bound. On page 346 of this issue, Wienholtz and collaborators[1] report precise measurements of the masses of two short-lived, neutron-rich calcium nuclei, calcium-53 and calcium-54, which until now researchers have been unable to weigh. By comparing the measured masses with state-of-the-art model calculations aimed at describing the properties of atomic nuclei at a microscopic level, the authors have confirmed the unique role of calcium nuclei in deciphering the ingredients of the nuclear force that binds protons and neutrons together.

The atomic nucleus is a multifaceted quantum system comprising particles known as quarks and gluons, which interact to form protons and neutrons bound by the strong and electroweak forces. It is the primary system in nature in which these two fundamental forces govern everyday behaviour. Of the 3,000 or so known nuclear species, which differ in their proton and neutron numbers, fewer than 300 are stable and occur naturally. The others tend to decay until a stable nucleus forms, often existing for only fractions of a second. The study of these exotic nuclei has proved crucial for our understanding of the complex interplay of constituents within a nucleus.

One of the overarching goals of nuclear physics is the development of a comprehensive model of the atomic nucleus that can predict the physical properties of all possible nuclei. The ability to predict the properties of extremely short-lived nuclei is essential if we are to understand the origin of the elements in the Universe. The most abundant elements, hydrogen and helium, were formed shortly after the Big Bang. Slightly heavier elements, such as carbon and oxygen, are constantly generated in the nuclear reactions that fuel stars. Elements heavier than iron are thought to be produced in extreme environments, with perhaps more than half originating in supernovae — the most violent stellar explosions.

Many of the nuclear reactions and decay sequences that underpin element-forming processes involve short-lived nuclei that are not normally found on Earth. Furthermore, some of the nuclei that participate in these processes cannot be studied in the laboratory. Nuclear astrophysicists, therefore, have to rely on models of atomic nuclei that pertain in the regime of short-lived nuclear species. Understanding the complex nuclear force is a necessary step in the construction of such models, and a formidable challenge. The precise mass measurements made for ^{53}Ca and ^{54}Ca by Wienholtz and colleagues provide a key benchmark for present and future nuclear models.

The masses of many nuclei can be measured precisely using Penning-trap mass spectrometry[2], in which a charged particle is trapped in a strong magnetic field so that it undergoes oscillatory 'cyclotron' motion. The frequency of this motion is proportional to the mass-to-charge ratio of the particle and to the magnetic field strength of the trap, thus allowing the particle's mass to be deduced. Relative uncertainties of less than 1 part in a billion have been obtained[3] for stable nuclei that can be confined individually in a trap for long periods — a few days in some cases. Analysing exotic nuclei is challenging because, aside from their limited half-lives, these nuclei are rarely available in great abundance and/or do not come as pure ensembles. Nevertheless, extremely short-lived nuclei (such as those that have half-lives of a few milliseconds) have been measured successfully using Penning traps[2].

Unfortunately for Wienholtz *et al.*, this approach was not possible for ^{53}Ca and ^{54}Ca. These isotopes are remarkably difficult to make, and so the authors needed to use the powerful ISOLDE facility at the CERN research centre near Geneva in Switzerland. ISOLDE produces rare nuclei from the proton-induced fission of uranium carbide. Even so, the production rates of the nuclei were low. For example, the authors detected only a few ^{54}Ca ions per minute, accompanied by copious amounts of the less-exotic contaminant chromium-54.

In a groundbreaking twist, the researchers used a newly installed system known as a multi-reflection time-of-flight mass spectrometer/separator (MR-TOF MS)[4] to measure the masses of ^{53}Ca and ^{54}Ca. This system speeds up the analysis, thus preventing the loss of fleetingly existent nuclei through radioactive decay before their masses can be measured in ISOLDE's Penning trap. In the MR-TOF MS, nuclei are sent rushing back and forth between electrostatic 'mirrors', allowing them to fly several kilometres within a region the size of a tabletop. The flight time of each particle is related to its mass-to-charge ratio and so, after their particles had travelled a sufficiently long flight path, Wienholtz *et al.* were able to separate the nuclei of interest from contaminants. Using this method, the authors determined the masses of ^{53}Ca and ^{54}Ca with uncertainties of less than 1 in 100,000. This allowed them to literally weigh up the effects of nuclear-constituent interactions.

Calcium is special because its proton number (20) corresponds to a particularly inert nuclear configuration that is equivalent to the electronic configurations of noble gases. This allows the effect of neutron richness on nuclei to be studied systematically. Furthermore, calcium's inert configuration means that the element marks the frontier for several of the theories that try to model the nuclear force from first principles. The measured masses of ^{53}Ca and ^{54}Ca, together with the masses of their less-exotic neighbours, ^{48}Ca to ^{52}Ca, provide a formidable benchmark for nuclear theory.

Indeed, Wienholtz and co-workers found that their experimental results compared exceptionally well to predictions made using their own theoretical approach[5]. Their model incorporates three-nucleon forces, an important component of the nuclear interaction[6], with parameters constrained from properties of light nuclei (such as the binding energy of the helium-3 nucleus, and the radius of helium-4), and provides a true prediction for the masses of medium-sized calcium nuclei.

Wienholtz and colleagues' experimental approach is highly promising as a method for making precise measurements of the masses of short-lived, exotic nuclei that would otherwise be impossible to achieve. It will inspire similar developments at nuclear-science facilities around the world. ■

Alexandra Gade *is at the National Superconducting Cyclotron Laboratory and the Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA.*
e-mail: gade@nscl.msu.edu

1. Wienholtz, F. *et al. Nature* **498,** 346–349 (2013).
2. Blaum, K. *et al. Phys. Scripta* **T152,** 014017 (2013).
3. Meyers, E. *Int. J. Mass Spectr.* http://dx.doi.org/10.1016/j.ijms.2013.03.018 (2013).
4. Wolf, R. N. *et al. Nucl. Instrum. Meth. A* **686,** 82–90 (2012).
5. Holt, J. *et al. J. Phys. G* **39,** 085111 (2012).
6. Hammer, H.-W. *et al. Rev. Mod. Phys.* **85,** 197–217 (2013).

STRUCTURAL BIOLOGY

Signalling from disordered proteins

The discovery that a disordered protein can transmit signals between two binding sites calls into question the idea that communication within proteins requires a specific structural pathway linking such sites. SEE LETTER P.390

VINCENT J. HILSER

Signalling is at the heart of biology, and most of it is mediated by proteins. However, despite recent progress in identifying the component proteins of many signalling networks[1], and in our understanding of the complexities of such networks, how individual proteins fulfil their role in a particular network is still not well understood. On page 390 of this issue, Ferreon *et al.*[2] show that intrinsically disordered proteins — which lack well-defined tertiary structures[3], unlike their structured protein counterparts — can not only regulate the magnitude of a particular signal, but also reverse it, transforming a positive effector into a negative one. This result reinforces an emerging view that such proteins play a crucial part in signalling[4], and challenges much of the current dogma about the relationship between protein structure and function, and

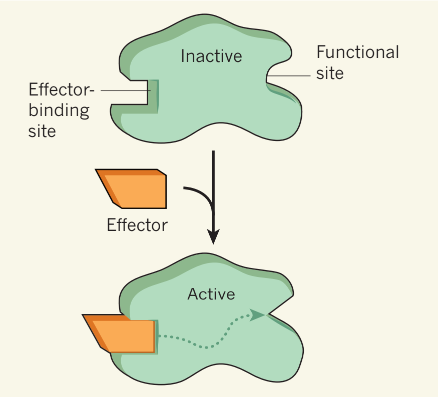


Figure 1 | The structural view of allostery. One model of allostery considers single protein molecules: the binding of an effector molecule makes, breaks and/or shifts bonds (dotted arrow) between the effector-binding site and the functional site of the protein, causing the functional site to adopt an active conformation.

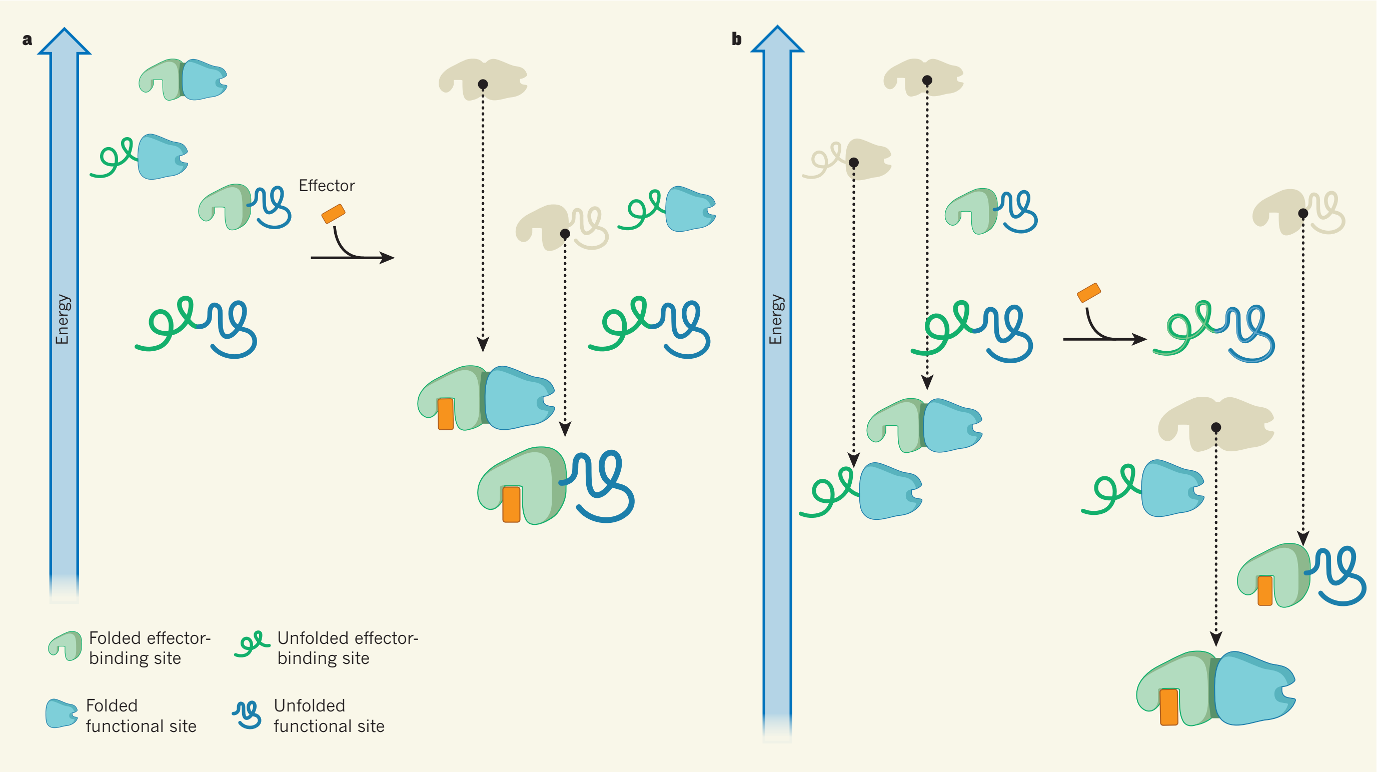


Figure 2 | The ensemble view of allostery. Ferreon and colleagues' results[2] support a model of allostery that considers ensembles of a protein molecule in different conformations. **a**, In the absence of an effector, protein conformations in which the functional site is folded (active) have high energies, and so the probability of the functional site being folded is low. Effector binding lowers the energies (dotted arrows) of protein conformations in which the effector-binding site is folded, including the conformation that contains the active form of the functional site; this increases the overall probability that the functional site will be active. The sizes of the depicted structures indicate the probability of their formation: larger structures are more likely to form. **b**, Perturbation of the protein in the absence of an effector alters the relative energies of different conformations, so that the two lowest-energy conformations have folded functional sites. Subsequent effector binding changes the relative energies so that only one of the two lowest-energy conformations has a folded functional site, lowering the overall probability that this site will be active.

the role that specific structural pathways have in signalling.

Activators and inhibitors are regulatory molecules (known as effectors) that respectively enhance or suppress the activity of proteins. In many cases, they work by binding to a protein at a site that is distant from the protein's functional site, presumably inducing a structural change that makes the functional site more compatible (activators) or less compatible (inhibitors) with its substrates — a process known as allostery. How these changes are mediated has been a central question in biology for more than 50 years.

The first allosteric protein to be crystallized was haemoglobin[5], which can bind four molecules of oxygen. It was known that the binding of one or two molecules to haemoglobin increases the protein's affinity for more oxygen molecules (that is, oxygen is a positive effector for subsequent oxygen binding), a phenomenon referred to as positive cooperativity. The observation of striking differences in the crystal structures of haemoglobin in the presence and absence of oxygen[6] seemed to validate the idea that allostery can be rationalized, and possibly even quantitatively accounted for, by examining the structural distortions that connect the different oxygen-binding sites.

This structural view of allostery (Fig. 1) has largely guided the field ever since. However, the realization that more than 30% of the proteome — the complete set of proteins found in a cell — consists of intrinsically disordered proteins (IDPs)[7], and that intrinsic disorder is hyper-abundant in allosteric signalling proteins such as transcription factors[8], raises the possibility that a well-defined structure is neither necessary nor sufficient for signal transmission.

So how do IDPs transmit signals? Using a spectroscopic technique that allows the distance between two sites on a single molecule to be determined, Ferreon *et al.*[2] directly measured the probability that a molecule of a viral IDP, called E1A, would form a complex with either one or both of the two host-encoded proteins (CBP and pRB) that it interacts with as it disrupts a host's cellular signalling. Importantly, E1A acquires ordered structure only when it is bound to its protein partners[9]. Earlier studies suggested that IDPs can use coupled folding and binding to facilitate allosteric coupling[10]. Ferreon and colleagues' results get to the heart of how this coupling is facilitated.

The idea underpinning the authors' work is that any protein exists as an ensemble of conformations. In the simplest case, an IDP has an effector-binding domain and a functional domain (Fig. 2a). Each of these two domains can be unfolded (inactive) or folded (active), and can be folded separately or together. In the absence of an effector molecule, the folded form of the effector-binding domain is unstable. Binding of the effector molecule stabilizes the folded form, which in turn can stabilize the folded form of the functional domain by redistributing the ensemble of conformations.

Ferreon and co-workers report that, for a long version of the E1A protein, the binding of CBP and pRB is positively coupled — the binding of E1A to either CBP or pRB increases the probability that it will bind to the other. Such positive coupling is abundant in nature and underlies almost all regulatory processes. But the authors observed something remarkable with a truncated version of E1A. Instead of positive coupling being wiped out or decreased, negative coupling occurred: the binding of either CBP or pRB lowered the probability that E1A would bind to the second protein.

This type of cooperativity switching is difficult to reconcile using a purely structural representation of signalling, because it arises from a balance of competing effects. As revealed by an ensemble representation of proteins, effector binding stabilizes both the active and inactive forms of the functional domain, which means that the effector is potentially an activator and an inhibitor. So what determines whether the effector will activate or inhibit?

The answer is the relative stability of each state in the ensemble. Under one set of conditions (Fig. 2a), the ensemble could be poised such that effector binding causes activation. But under another set (Fig. 2b), effector binding can cause inhibition. Crucially, a switch in cooperativity can arise as a result of any type of perturbation (such as the binding of another molecule, post-translational modification or protein truncation) that can redistribute the ensemble of conformations[11], even to the extent of transforming effector binding from activating to inhibiting, or vice versa.

Although Ferreon and colleagues' work does not reveal how the observed cooperativity switch occurs, it does help to clarify the following key questions that underlie a quantitative understanding of signalling in IDPs, and perhaps also in structured proteins. What states comprise the protein ensemble, and what are their probabilities? And are there ground rules that dictate whether signalling, or even activation–inhibition switching, can occur in an ensemble[10,11]? The take-home message of Ferreon and colleagues' work, and the reason that a switch is possible, is that proteins should not be thought of as multiple copies of identical structures that respond uniformly to a signal. Instead, proteins — especially IDPs — exist as ensembles of sometimes radically different structural states. This structural heterogeneity can produce ensembles that are functionally 'pluripotent', a property that endows IDPs with a unique repertoire of regulatory strategies. ■

Vincent J. Hilser *is in the Departments of Biology and Biophysics, Johns Hopkins University, Baltimore, Maryland 21218, USA.*
e-mail: hilser@jhu.edu

1. Hao, N., Budnik, B. A., Gunawardena, J. & O'Shea, E. K. *Science* **339,** 460–464 (2013).
2. Ferreon, A. C. M., Ferreon, J. C., Wright, P. E. & Deniz, A. A. *Nature* **498,** 390–394 (2013).
3. Wright, P. E. & Dyson, J. H. *J. Mol. Biol.* **293,** 321–331 (1999).
4. Xie, H. *et al. J. Proteome Res.* **6,** 1882–1898 (2007).
5. Perutz, M. F. *et al. Nature* **185,** 416–422 (1960).
6. Dickerson, R. E. *Annu. Rev. Biophys. Chem.* **41,** 815–842 (1972).
7. Ward, J., Sodhi, J., McGuffin, L., Buxton, B. & Jones, D. *J. Mol. Biol.* **337,** 635–645 (2004).
8. Liu, J. *et al. Biochemistry* **45,** 6873–6888 (2006).
9. Wright, P. E. & Dyson, J. H. *Curr. Opin. Struct. Biol.* **19,** 31–38 (2009).
10. Hilser, V. J. & Thompson, E. B. *Proc. Natl Acad. Sci. USA* **104,** 8311–8315 (2007).
11. Motlagh, H. & Hilser, V. J. *Proc. Natl Acad. Sci. USA* **109,** 4134–4139 (2012).

NATURE

50 Years Ago

By and large the effect of automization is to reduce severely the demand for unskilled and semi-skilled workers and to increase sharply the need for skilled workers … This trend is surely to be welcomed. Repetition work is an insult to the people who have to do it. It treats them as less than human. It is not surprising if it often turns them into something less than human. If you make a man spend eight hours a day in which he has nothing to … exercise his mental powers on, is it surprising that he is incapable of exercising those powers in his leisure time and must spend it watching television or wrecking a dance hall? Automation offers the prospect of giving every man and woman a job that is interesting and worth doing in itself, a job requiring initiative or creative thought. Surely that is as desirable an object as providing a higher standard of material living.
From *Nature* 22 June 1963

100 Years Ago

After expressing his admiration for the character of Wilbur Wright … the lecturer considered the resemblance and differences of the manufactured aëroplane and the living bird. The resemblance may be simply the result of copying the bird, or it may be that similar designs have been arrived at independently by birds and men … These resemblances are remarkable, but there are great differences … No flying animal uses a continuously rotating propeller to drive him forward on soaring wings, and it is perhaps scarcely too much to say that if birds only knew how, they would now copy the Wright brothers. Muscular action and the circulation of the blood, however, put supreme difficulties in the way of the development of the continuous rotation of a part of an animal.
From *Nature* 19 June 1913

VIROLOGY

The virus whose family expanded

The discovery of many new species of hepaciviruses and pegiviruses, which exhibit enormous genetic diversity, in wild rodent and bat populations might help us to understand the origins of the hepatitis C virus.

OLIVER G. PYBUS & REBECCA R. GRAY

The hepatitis C virus does not give up its secrets lightly. Despite infecting about 3 out of every 100 people worldwide, a small proportion of whom consequently develop severe liver disease, the virus eluded discovery for decades. It was eventually identified in 1989 as the cause of 'non-A, non-B hepatitis'. Researchers who have since sought the origins of hepatitis C virus (HCV), as it is now known, have been frustrated in equal measure. The virus infects chimpanzees in the laboratory, but studies of wild and captive primates uncovered no evidence of an animal population that might have transmitted HCV to humans[1], contrasting with the success of other surveys that exposed close relatives of HIV-1 and human malaria in great apes[2]. Now, however, Kapoor *et al.*[3] and Quan *et al.*[4], writing in *mBio* and *Proceedings of the National Academy of Sciences*, respectively, report a diverse and widespread array of HCV-like viruses in wild populations of rodents[3] and bats[4]. Although none of these viruses can yet be claimed as the source of HCV, their discovery may represent the beginning of the end of the search for HCV's origins.

HCV belongs to the *Hepacivirus* genus of viruses, whose closest taxonomic neighbour is the *Pegivirus* genus[5]; the newly discovered bat and rodent viruses include members of both groups. Kapoor *et al.* found five provisional virus species among more than 400 blood samples from four North American rodent species. Quan and colleagues describe 11 virus lineages

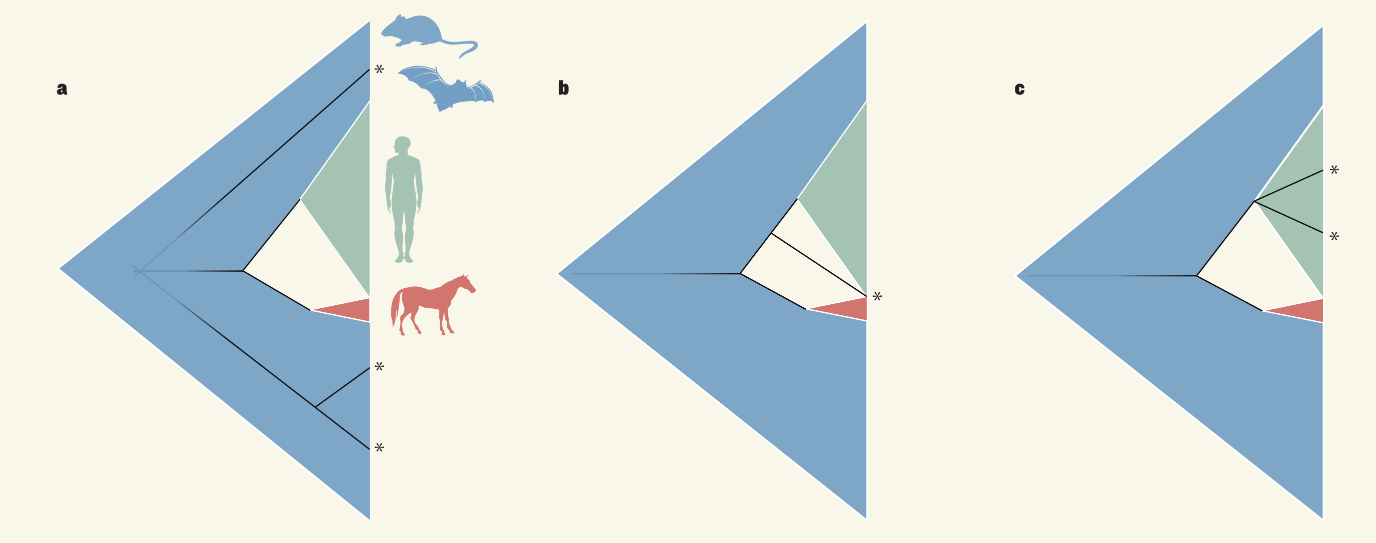


Figure 1 | Possible evolutionary trees of the hepaciviruses. Triangles represent the large genetic diversity of the hepaciviruses discovered by Kapoor *et al.*[3] and Quan *et al.*[4] in bats and rodents (blue), and the more limited diversity of human hepatitis C viruses (HCV; green) and the hepaciviruses found in horses (red). Future surveys in bats, rodents or other animals may discover more hepaciviruses (asterisks), the evolutionary position of which would define three possible scenarios for the origins of HCV. **a**, None of the new viruses is closely related to HCV and its origin remains unresolved. **b**, Viruses more similar to HCV than to equine hepacivirus, HCV's closest known relative, are found. This would suggest that all HCV strains arose from a single ancestral transfer to humans. **c**, The new viruses group within the known genetic diversity of HCV, indicating that it arose from two or more independent cross-species transmissions.

from around 1,700 samples taken from 58 bat species collected in Mexico, Bangladesh and sub-Saharan Africa. The most notable property of the new viruses is their exceptional genetic heterogeneity, which dwarfs the diversity of all previously known hepaciviruses and pegiviruses, including HCV, which is itself highly variable.

This diversity strongly implicates bats and rodents as natural and ancestral hosts for viruses of both genera, an idea supported by the comparatively high frequency of infection in wild animals (around 5%) and by Quan and colleagues' observations that some bats were co-infected with multiple viruses. Furthermore, all the infected bats seemed healthy when collected, which is consistent with a long evolutionary association between virus and host. But despite their already remarkable diversity, the viruses were isolated from approximately 5% of bat and less than 1% of rodent species known, and thus probably represent only a fraction of hepaciviruses and pegiviruses present in nature.

Before these reports, the hepaciviruses and pegiviruses were known as sparsely populated genera that between them contained fewer than ten species, isolated from a motley collection of hosts: humans, chimpanzees, horses, dogs, wild and captive New World primates, plus one bat pegivirus found[6] in 2010. The discovery of enormous viral genetic diversity in bats and rodents presents the possibility that each of the formerly identified species arose through successful cross-species transmission of a bat or rodent virus. Indeed, it is estimated that a quarter of recently emerged human pathogens originated from rodents or bats[7], and both animal groups are abundant, widely distributed and live in large numbers near human settlements or domesticated animals. This postulated cross-species transmission need not have been direct, but may have occurred through an intermediate host in even closer contact with humans — civet cats had such a role in the transfer of the SARS coronavirus to humans[8], and pigs in the transfer of the Nipah virus[9], both of which originate in bats.

Although none of the new hepaciviruses and pegiviruses are sufficiently genetically similar to those found in humans or other animals to be declared their immediate source, bats and rodents are now prime suspects in the hunt for the ultimate origins of HCV. Further sampling of small-mammal populations worldwide should reveal the true diversity and host range of these viruses, and may uncover viruses more similar to HCV. Three possible outcomes of such sampling can be imagined: new viruses are found but none are closely related to HCV and its origin remains unresolved (Fig. 1a); viruses more similar to HCV than to equine hepacivirus, HCV's closest known relative, are found, suggesting that all HCV strains arose from a single successful ancestral transfer to humans (Fig. 1b); or viruses are found that group within the current genetic diversity of HCV, indicating that it arose from two or more independent cross-species transmissions (Fig. 1c).

The third hypothesis is particularly intriguing as it potentially solves the enigma of 'endemic' HCV transmission: how some rural populations in central Africa and southeast Asia come to bear a range of divergent HCV strains, indicative of centuries of stable human-to-human transmission, in the absence of any consistently effective and widespread route of transmission. This riddle would be answered if the virus diversity originates not in humans but from an animal reservoir.

Although the immediate consequences of the current findings for human health seem minimal, only detailed investigation of the transmission and ecology of the new viruses in their natural hosts can elucidate their true potential for cross-species transmission. The ongoing emergence in humans of coronaviruses of probable bat origin[10], ten years after the successful eradication of SARS, is a timely reminder of the potential benefits to epidemiology and public health of understanding the dynamics of infectious disease in wild animal populations. ■

Oliver G. Pybus *and* **Rebecca R. Gray** *are in the Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.*
e-mail: oliver.pybus@zoo.ox.ac.uk

1. Makuwa, M. *et al. J. Med. Primatol.* **35,** 384–387 (2006).
2. Sharp, P. M., Rayner, J. C. & Hahn, B. H. *Science* **340,** 284–286 (2013).
3. Kapoor, A. *et al. mBio* **4,** e00216-13 (2013).
4. Quan, P. L. *et al. Proc. Natl Acad. Sci. USA* **110,** 8194–8199 (2013).
5. Stapleton, J. T., Foung, S., Muerhoff, A. S., Bukh, J. & Simmonds, P. *J. Gen. Virol.* **92,** 233–246 (2011).
6. Epstein, J. H. *et al. PLoS Pathogens* **6,** e1000972 (2010).
7. Woolhouse, M. & Gaunt, E. *Crit. Rev. Microbiol.* **33,** 231–242 (2007).
8. Li, W. *et al. Science* **310,** 676–679 (2005).
9. Chua, K. B. *et al. Science* **288,** 1432–1435 (2000).
10. van Boheemena, S. *et al. mBio* **3,** e00473-12 (2012).

doi:10.1038/nature12217

Anisotropic leaky-mode modulator for holographic video displays

D. E. Smalley[1], Q. Y. J. Smithwick[1], V. M. Bove Jr[1], J. Barabas[1] & S. Jolly[1]

Every holographic video display is built on a spatial light modulator, which directs light by diffraction to form points in three-dimensional space. The modulators currently used for holographic video displays are challenging to use for several reasons: they have relatively low bandwidth, high cost, low diffraction angle, poor scalability, and the presence of quantization noise, unwanted diffractive orders and zero-order light. Here we present modulators for holographic video displays based on anisotropic leaky-mode couplers, which have the potential to address all of these challenges. These modulators can be fabricated simply, monolithically and at low cost. Additionally, these modulators are capable of new functionalities, such as wavelength division multiplexing for colour display. We demonstrate three enabling properties of particular interest—polarization rotation, enlarged angular diffraction, and frequency domain colour filtering—and suggest that this technology can be used as a platform for low-cost, high-performance holographic video displays.

The limitations and useful properties (affordances) of holographic video displays are chiefly dictated by the spatial light modulators (SLMs) on which they are built. The temporal bandwidth of the spatial light modulator determines the display size, view angle and frame rate. The pixel pitch determines the angle of the display or the power of the lenses needed to achieve a wide view angle. The space–bandwidth product, which is related to the numerical aperture of the holographic grating, determines the maximum depth range and number of resolvable views the display will possess. Finally, the optical non-idealities of the modulator give rise to noise and artefacts in the display output. Current state-of-the-art technologies for spatial light modulation (for example, liquid crystal, micro-electro-mechanical systems (MEMS)[1,2], and bulk-wave acousto-optic modulators[3]) have proven challenging to employ in holographic video displays. Before using these modulators in a holographic display, one must address their low bandwidth, low diffraction angle, quantization error, and the presence of zero order and other noise (see Fig. 1) as well as the spatial or temporal multiplexing of colour. Much of the cost and complexity of modern holographic displays is due to efforts to compensate for these deficiencies by, for example, adding eye tracking to deal with low diffraction angle[4], duplicating and phase shifting the optical path to eliminate the zero order[5], or creating large arrays of spatial light modulators to increase the display size[6]. The cost and complexity of holographic video displays could be greatly reduced if a spatial light modulator could be made to have better affordances than the liquid crystal and MEMS devices currently used.

We have developed a spatial light modulator based on anisotropic leaky-mode coupling that brings the tools of guided wave optics to bear on the challenges of holographic video and possesses many advantages over liquid crystal and MEMS devices when applied to holographic video display. Here we describe how the device can be fabricated inexpensively and made to support an aggregate temporal bandwidth of more than 50 billion pixels per second (50 Gpixels s^{-1}) —an order of magnitude increase over the current state-of-the-art. (A graphical representation of the modulator fabrication process can be found in Supplementary Figs 1 and 2.) We also demonstrate a threefold increase in angular deflection over other modulator technologies due to the edge-lit nature of the waveguide grating structure and the resulting increase in space–bandwidth product. The modulator exploits guided-wave phenomena, most notably anisotropic mode conversion for the elimination of zero-order light and tunable wavelength filtering for the simultaneous and superimposed modulation of colour signals.

Structurally, an anisotropic leaky-mode coupler is a proton-exchanged[7] channel waveguide on a lithium niobate substrate with a transducer at one end[8,9]. The waveguide is anisotropic and only guides light in one polarization. When excited by a radio frequency signal, the transducer generates surface acoustic waves[10] (SAWS) that propagate collinearly with the light trapped in the anisotropic waveguide (see Fig. 2a). When the phase-matching condition is met,

$$\beta_{\text{guided}} - K_{\text{grating}} = \beta_{\text{leaky}} \tag{1}$$

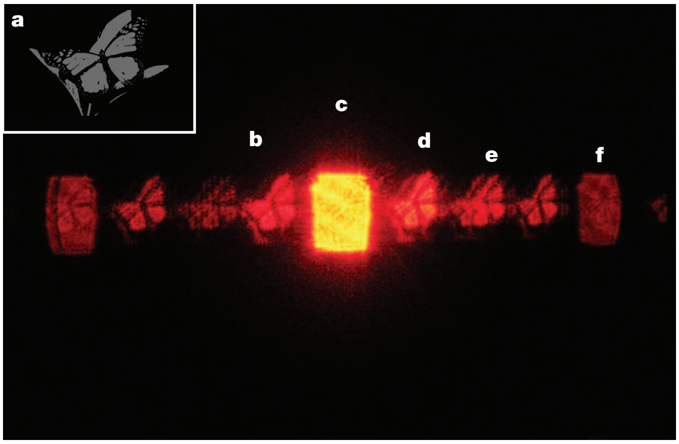


Figure 1 | Artefacts from a holographic stereogram on a pixelated (liquid crystal on silicon) modulator. **a**, The stereogram mask; **b**, intended output; **c**, zero order (undiffracted light); **d**, unwanted conjugate image; **e**, higher-order images and quantization noise; **f**, diffracted order arising from the modulator pixel structure. The scene used to generate this stereogram was provided by J. Buchholz.

[1] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

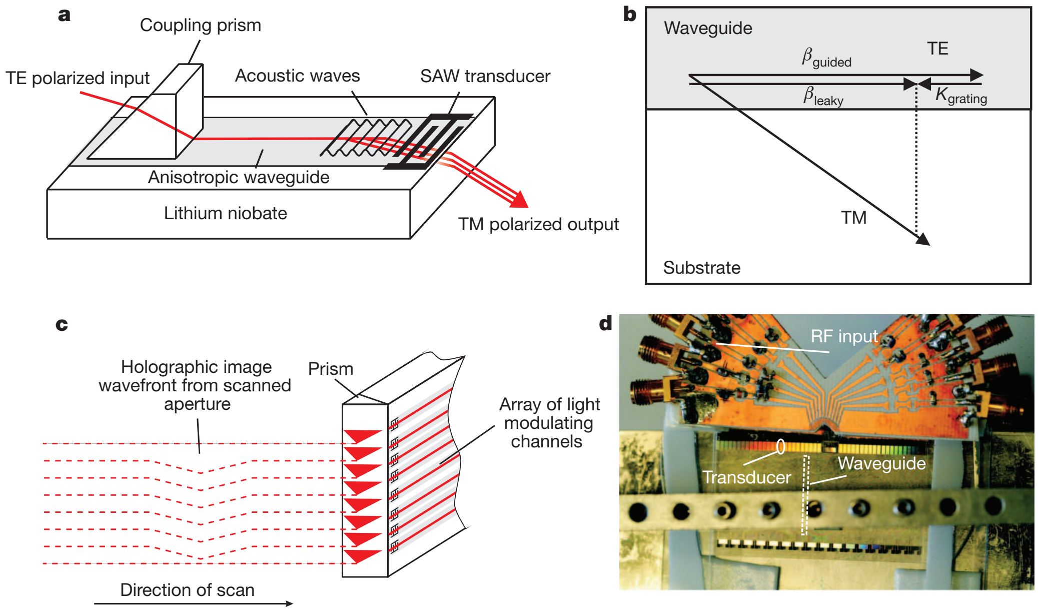


Figure 2 | The structure and function of anisotropic mode-coupling modulators and modulator arrays. **a**, Single channel anisotropic mode-coupling modulator. Guided, TE polarized light is converted by acoustic waves (launched from an SAW transducer) into leaky TM polarized light. The acoustic waves act as the holographic diffraction pattern. **b**, Phase matching condition for mode coupling. **c**, The holographic image is formed by scanning the aperture of the anisotropic waveguide device having one or more channels. **d**, A multichannel anisotropic waveguide modulator. The modulator pictured has more than 40 channels. Devices with as many as 1,250 channels are being fabricated.

where β_{guided} is the wavevector of the guided TE (transverse electric) mode, $K_{grating}$ is the grating vector corresponding to the acoustic pattern encoded with holographic information, and β_{leaky} is the component of the wavevector of the leaky TM (transverse magnetic) mode along the direction of the grating vector and the guided mode (see Fig. 2b).

The acoustic pattern, encoded with holographic information, couples the guided light into a leaky mode of orthogonal polarization which leaves the waveguide–substrate interface. The index contrast of the waveguide–air interface is much higher than that of the waveguide–substrate interface; this asymmetry of boundary conditions means that there is no conjugate image (an unwanted mirror image of the hologram output that is formed by symmetric gratings). This leaky mode emits a wavefront-modulated fan of light that leaves one face of the wafer and forms part of a holographic output image. Each channel waveguide writes one or more lines of the output, and several channels can be fabricated next to each other to create large aggregate bandwidths suitable for large display size and resolution (see Fig. 2c). Such a fabricated multichannel device is shown in Fig. 2d.

Advantages of leaky-mode couplers

Anisotropic leaky-mode couplers possess several advantages over other spatial light modulators used for holographic video (see Table 1). In addition to being simple to fabricate and drive, they are capable of high deflection for a given spatial grating pitch and can make use of tools from guided-wave optics to address noise and colour multiplexing.

Modulators with defined pixel structure and a backplane (for example, liquid crystal and MEMS devices) become more complex as pixels are added, which constrains scalability. Bulk-wave acousto-optic modulators can produce the acoustic equivalent of 100 million pixels per second (100 Mpixels s^{-1}) per acoustic channel; however, channels cannot be placed too closely together because of the resulting crosstalk. Anisotropic leaky-mode couplers enjoy lateral guidance of the acoustic wave, which makes it possible for adjacent channels to be placed tens of micrometres apart and for hundreds of channels to be placed side-by-side on a single substrate, thereby providing aggregate bandwidths in excess of 50 Gpixels s^{-1}. This bandwidth is nearly an order of magnitude greater than the temporal bandwidth of current pixelated modulators. A device with 500 channels could provide enough bandwidth to drive a horizontal-parallax only (HPO) holographic display one metre in width. At the time of publication, we are fabricating devices with as many as 1,250 channels.

Fabrication of active liquid crystal and MEMS devices requires as many as 20 or more mask steps to define both the pixels and the associated backplane. Only two masks are required to fabricate guided-wave modulators: one to define the waveguide structure and one to pattern the transducers. The resulting fabrication and cost are similar to that of common SAW filters which sell for a dollar or less. A device capable of producing standard resolution HPO holographic video images would cost in the low tens of dollars to fabricate, as a conservative estimate.

Guided-wave modulators are analogue devices and can be driven by up-converted, standard analogue video signals, generated by, for

Table 1 | Advantages of anisotropic waveguide modulators

Property	Pixelated modulator	Anisotropic waveguide modulator
Temporal bandwidth	5 Gpixels s^{-1} (assuming an 8 Mpixel SLM)	50 Gpixels s^{-1} (assuming a 500 channel modulator)
Output angle ($\lambda = 532$ nm, $\Lambda = 12$ μm)	2.54°	24.7°
Output polarization orthogonal to zero order?	No	Yes
Superfluous orders at output	Multiple	None
Fabrication complexity	20 masks	2 masks
Superfluous conjugate mode	Yes	No
Hologram approximation basis	Quantized pixels	Sinusoidal waves
Colour multiplexing	Space/time	Space/time/frequency

Pixelated modulators considered here are MEMS and liquid crystal devices. All values are approximate. It should be noted that the angle of the output light in an anisotropic modulator is a function of waveguide parameters, such as the orientation of the substrate material (lithium niobate, *x*-cut, *y*-propagating in this case), and the wavenumber of the guided mode.

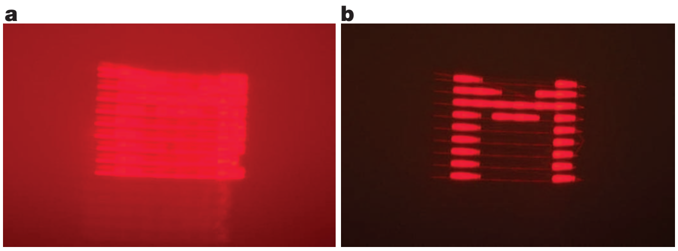


Figure 3 | Polarization rotation to exclude noise. **a**, **b**, The scanned output of the modulator is shown without a polarizer (**a**) and with a polarizer to exclude noise (**b**).

example, standard graphics cards commonly used in high-end graphics work. Because the modulators are analogue and have no pre-defined pixel microstructure, there is no intrinsic quantization of the signal. The device transducers can be used as filters to band-limit quantization noise that might be present in the video signal. As with pixelated modulators, light may diffract from harmonics of the acoustic signal, giving rise to higher-order diffracted signals; however, in anisotropic mode couplers, typically only one order is present at the output of the device. This is because conjugate modes are prohibited by waveguide asymmetry and higher-order modes are suppressed at the output by high angular separation of orders and total internal reflection.

In addition to the points given above, we elaborate here on three advantages of particular interest made possible by the waveguide nature of the device: hologram polarization rotation, increased angular deflection, and simultaneous and superimposed red-green-blue (RGB) modulation.

Polarization rotation

The waveguide in the guided-wave acousto-optic modulator is anisotropic so that it supports guided modes of only one polarization; modes of the orthogonal polarization are leaky. The acoustic signal couples light from the fundamental extraordinary guided mode to the first order leaky mode, rotating its polarization along the way[11,12]. As a result, the holographic image produced by the anisotropic waveguide modulator has a polarization that is orthogonal to all of the other light in the system. This allows noise, including zero-order light, to be excluded from the output with a polarizer, as shown in Fig. 3.

Wider angular deflection

Because the acoustic wave is being effectively illuminated by light at a glancing angle rather than at normal incidence, the resulting diffracted angle can be more than three times higher than it would be at normal incidence on another modulator of the same pixel pitch. This is shown in Fig. 4a, which was generated from the grating equation

$$\sin\theta_{\mathrm{out}} - \sin\theta_{\mathrm{in}} = \frac{m\lambda}{\Lambda} \qquad (2)$$

where θ_{in} is the angle of the illumination light, θ_{out} is the angle of the output light, Λ is the grating period, λ is the wavelength of light used, and m is the diffracted order. Standard modulators are illuminated near the grating normal but waveguided light interacts with the acoustic grating nearly collinearly. The differential effect of output angle with incident angle is shown in Fig. 4b. This effect is further magnified when the grating is inside a high-index material, as is the case in waveguide modulators. This is because the signal light is further deflected by refraction at the output face of the substrate. For the anisotropic modulator demonstrated here, the output angle for 532 nm light was measured to be 24.7° for a 12 μm period acoustic grating generated on the device by a 326 MHz radio-frequency signal. Because the anisotropic interaction limits the usable bandwidth of the modulator to approximately 50 MHz per colour (ref. 8), and because we use demagnification in our supporting optics to choose the final display view angle, only a fraction (2.6° for 532 nm light) of this angular extent is used. The modulator will present an output that, when scanned, looks like a 1 m image with a 2.6° view-zone. This image will be demagnified for a final display output with approximately 10 cm of extent and a 26° view-zone. Having a small input angle and large demagnification ratio is intentional in our display, as it reduces the requirements placed on the scanning optics and keeps the display compact. In our display geometry, the chief advantage of this angular expansion in anisotropic devices is that it gives approximately a fivefold increase in the rate of angular deflection (degrees of deflection per MHz of signal bandwidth) than is typically available to lithium niobate acousto-optic deflectors, bringing the angular rate of deflection of the anisotropic modulator almost to parity with slow shear mode tellurium dioxide Bragg cells but at a fraction of the cost and with the added advantages of lower acoustic attenuation and dramatically higher channel capacity.

Simultaneous, superimposed RGB modulation

Anisotropic waveguide devices are capable of multiplexing colour in frequency rather than in time or space. In liquid crystal, MEMS and bulk wave acousto-optic modulators, it is necessary either to dedicate pixels to one colour or to illuminate the SLM sequentially, thereby reducing the resolution or the maximum refresh rate. However, waveguide devices can use wavelength division multiplexing, which allows

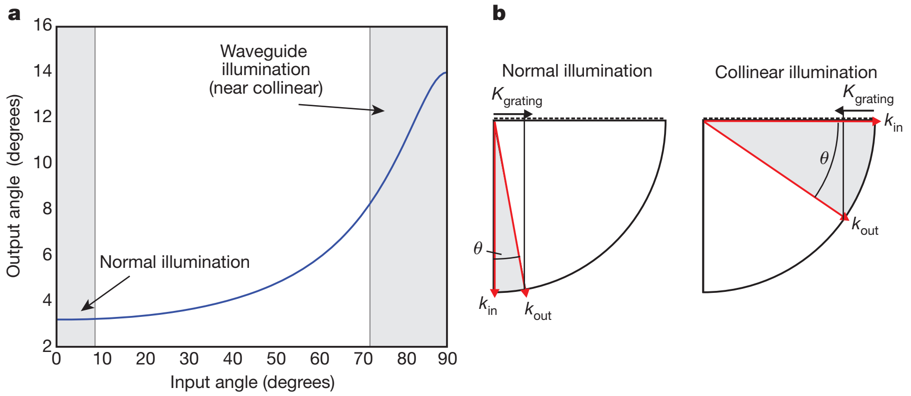


Figure 4 | Waveguide illumination for larger angular diffraction. **a**, Diffraction output angle versus input illumination angle for a 10 μm period grating illuminated with 633 nm light. Pixelated modulators are illuminated at angles near the perpendicular (these near-perpendicular angles are indicated by the left-most grey region), which affords a smaller range of diffracted output angles than is possible for a device illuminated at nearly collinear angles (near-collinear angles are indicated by the right-most grey region) as is the case in our anisotropic waveguide modulator. **b**, Angular output magnification for near-collinear waveguide illumination (right) relative to illumination at normal incidence (left) where k_{in} is the momentum vector of the input light (the illumination), k_{out} is the momentum vector of the diffracted output light, θ is the diffraction angle (highlighted by grey regions) and K_{grating} is the momentum vector of the grating. Note that θ is much larger for collinear illumination even though K_{grating} is the same in both cases.

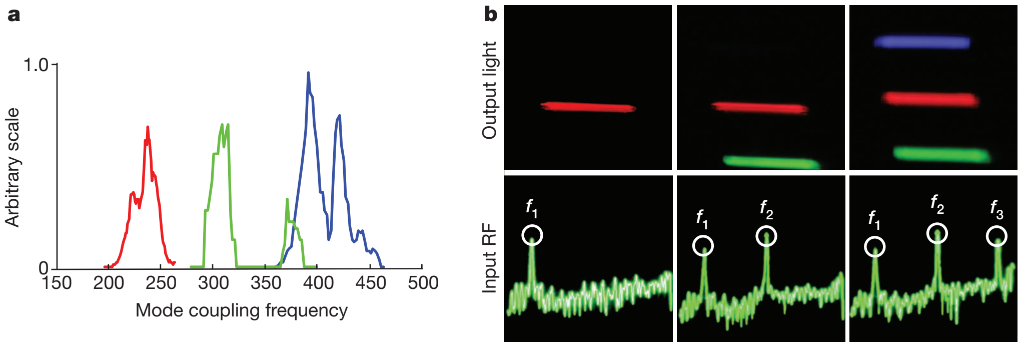


Figure 5 | Wavelength division multiplexing for colour displays. **a**, Frequency response of the anisotropic mode coupling device for red, green and blue light. **b**, Frequency multiplexing of red, green and blue light. The left panels show red output light for a low frequency input, f_1. The middle panels show both red and green output for an input containing both low frequency, f_1, and middle frequency, f_2, information. The right panels show red, green and blue outputs for an input signal containing low, medium and high frequencies (f_1, f_2 and f_3 respectively).

for simultaneous and superimposed modulation of red, green and blue light, so no colour filter wheel or separation of red, green and blue channels is necessary. This effect arises because the phase matching condition is wavelength-dependent. Red light mode converts at a lower frequency than green light, which in turn couples at a lower frequency than blue, allowing one to choose which colour to modulate by 'colouring' the frequency spectrum of the electrical signal sent to the modulator's transducers (see Fig. 5a). Because each channel is essentially a white-light emitter, the illumination of the device becomes trivial. Each channel or group of channels can be flood-illuminated by continuous red, green and blue light sources. This interaction is particularly well suited for colour holographic displays because the phenomenon of leaky mode coupling allows enough bandwidth for each colour to scan out a useful fan of angles but at the same time each passband is sufficiently separated to allow for independent operation. Additionally, it is also very convenient that all three colour bands fit approximately within the 200 MHz available from analogue video outputs of standard graphics processors.

To demonstrate simultaneous, superimposed RGB modulation, we illuminated one channel of an anisotropic waveguide array with continuous red, green and blue light ($\lambda = 633$ nm, $\lambda = 532$ nm, and $\lambda = 445$ nm). We stimulated a single, wideband transducer with a radio-frequency signal containing colour information that was separated in frequency with red information centred at 213 MHz, green at 333 MHz and blue at 387 MHz. The diffracted output of the modulator was scanned with x–y galvanometric mirrors to generate the test pattern in Fig. 5b. Then the output of the modulator was de-scanned with a rotating polygon and multiplexed vertically with a galvanometer to generate the holographic stereogram images (see Fig. 6) using a modified Scophony architecture[3,13] (see Fig. 7). The holographic stereograms were displayed at a resolution of 156 pixels × 177,600 pixels and at a refresh rate of 5 frames s^{-1} (here frame rate was traded for vertical resolution so an image could be made from a single channel device).

Other modulator parameters

Other parameters important to spatial light modulators are diffraction efficiency, temporal bandwidth, space–bandwidth product and cost. Our devices have had a wide range of efficiencies, all less than 10% for 0.5 W of applied radio-frequency power. Several other researchers, with more optimized designs (better-quality waveguides, narrower channels and carefully tuned annealing times), have reported efficiencies up to 90% (for 0.58 W of applied power) with room for additional improvement[14–16]. The 3 db bandwidth per channel of the device used here has been measured to be approximately 40 MHz per colour when using a uniform transducer(see Fig. 5a) and approximately 60 MHz when using a chirped transducer. This is consistent with the literature[8]. Given an acoustic wave speed of approximately 3,700 m s^{-1} and taking 50 MHz as the channel bandwidth, the space–bandwidth product of the device aperture is 13.5 cycles per millimetre of interaction length. The interaction length can be as great as 50 mm if limited only by acoustic attenuation. The maximum number of cycles in the scanned aperture of a 30 Hz display is 1.67 Mpixels per channel per frame. The device used here, fabricated as part of a two-wafer run which included electron-beam lithography, cost approximately US$50 to process at MIT's fabrication facilities. By processing 10 wafers at a time and by using photolithography, rather than electron-beam direct writing to define the transducers, the same device could be fabricated for less than US$3. Furthermore, an inexpensive geometry suitable for housing this device has been described[17].

Considerations for displays

Given the advantages described here, a new family of flexible holographic video displays is now possible. In holographic video displays using anisotropic mode couplers, the output of the device is scanned to create large outputs by persistence of vision. Because the modulator is an analogue device, display parameters such as frame rate, view angle, image extent and vertical resolution can be interchanged fluidly as long as the bandwidth budget is satisfied. If more space–bandwidth product (which is related to the concept of numerical aperture and to the total number of scannable points in diffractive systems) is needed, the length of the channels can be extended to provide longer interaction lengths in accordance with the expression $N = L(\Delta f/v)$, where N is the space–bandwidth product (or number of scannable points), L is the channel length, v is the velocity of the acoustic wave and Δf is the bandwidth of the anisotropic mode coupling interaction. If more temporal bandwidth is needed, more channels can be added to the

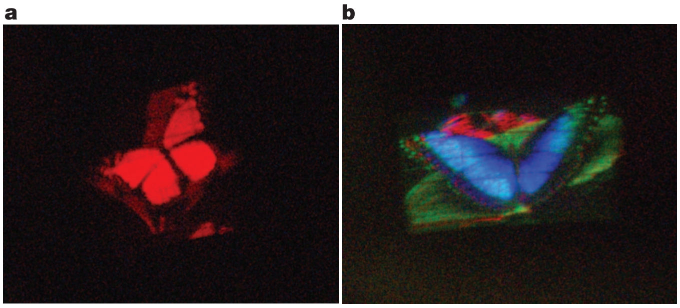


Figure 6 | Holographic stereograms made with a single channel anisotropic waveguide modulator, measuring 35 mm by 20 mm at the output of the display. **a**, Monochrome holographic stereogram. **b**, Colour holographic stereogram using simultaneous and superimposed modulation of red, green and blue light. The scene used to generate the stereogram in **a** was provided by J. Buchholz; Neil Doren Photography provided the live scene used to generate the holographic stereogram in **b**.

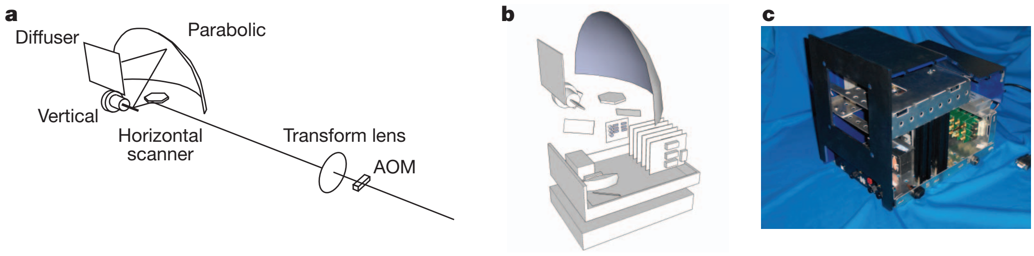


Figure 7 | PC-driven holographic video monitor. **a**, Holographic video monitor optical path containing a modulator, transform lens, horizontal polygon scanner, vertical galvanometric scanner, parabolic output lens and diffuser. **b**, Internal path folding of holographic video monitor. **c**, Assembled holographic video monitor.

modulator. When there are enough channels in an array to write all the necessary output lines simultaneously, there is no longer a need for vertical scanning and the problem of holographic video display becomes reduced to a single axis scan. With all lines written at once, the scanning optics are only required to make a full sweep once every 1/30th or 1/60th of a second, greatly expanding the size and type of scanning elements that may be used, which, interestingly, means that large displays can be more physically parsimonious than small ones.

Having demonstrated the advantages of anisotropic mode couplers, we are now exploring displays based on arrays of these devices such as a small, PC-driven, holographic video monitor and large-scale displays exceeding half a metre in width driven by dedicated hardware. Given the recent progress made in using graphics processing units (GPUs) for hologram fringe computation[18–20], it is now possible, using anisotropic mode coupling arrays driven by a commodity PC with a bank of high-end graphics cards, to make holographic video monitors with full-colour, standard video resolution and a 30 Hz refresh rate. Our research shows such a monitor might be constructed for less than US$500 (not including light sources). We are also investigating dedicated hardware solutions for driving large displays requiring tens of gigapixels per second.

METHODS SUMMARY

The modulators used here were fabricated from wafers of *x*-cut lithium niobate. The waveguides were formed by annealed proton exchange. The waveguides were defined by contact lithography. The transducers were defined by either contact lithography or direct electron-beam writing. The devices were impedance matched with lumped L-networks. Light was coupled into the waveguides using a rutile prism. The holographic stereogram images were created by taking one stereogram view at a resolution of 296 pixels × 156 pixels, stretching its resolution to 29,600 pixels × 156 pixels, and finally stitching 12 of these images together for a composite resolution of 355,200 pixels × 156 pixels. The α values of each of the red, green and blue channels of this image were multiplied by a different sinusoidal pattern in an OpenGL shader. All three colour signals were summed and divided by three, and sent out one of the video card outputs (for example, the nominal 'red' channel). This signal was then up-converted and amplified before entering a single transducer of the modulator array. Light from three lasers (at $\lambda = 445$ nm, $\lambda = 532$ nm and $\lambda = 633$ nm) was combined in an X cube and focused with an achromatic lens into one channel of an anisotropic leaky-mode coupling array. The output of the device was spatially filtered and focused on to the face of a spinning polygon (to optically descan the holographic fringe pattern so that it would appear stationary), vertically scanned onto a parabolic mirror (using the geometry shown in Fig. 7a), and finally imaged by a camera. For simplicity, only the view entering the camera was computed and displayed. The vertical diffuser shown in Fig. 6a, which extends the vertical viewzone of HPO holograms, was not used. A graphical representation of the modulator fabrication process can be found in Supplementary Figs 1 and 2.

Full Methods and any associated references are available in the online version of the paper.

Received 30 November 2012; accepted 23 April 2013.

1. Kreis, T., Aswendt, P. & Hofling, R. Hologram reconstruction using a digital micromirror device. *Opt. Eng.* **40,** 926–933 (2001).
2. Pearson, E. *MEMS Spatial Light Modulator for Holographic Displays*. Masters thesis, Massachusetts Institute of Technology (2001).
3. Hilaire, P., Benton, S. & Lucente, M. Synthetic aperture holography: a novel approach to three-dimensional displays. *J. Opt. Soc. Am. A* **9,** 1969–1977 (1992).
4. Häussler, R., Schwerdtner, A. & Leister, N. Large holographic displays as an alternative to stereoscopic displays. *Proc. SPIE Stereosc. Displays Applicat.* **XIX,** 68030M (2008).
5. Chen, G.-L., Lin, C.-Y., Kuo, M.-K. & Chang, C.-C. Numerical suppression of zero-order image in digital holography. *Opt. Express* **15,** 8851–8856 (2007).
6. Sato, K., Sugita, A., Morimoto, M. & Fujii, K. Reconstruction of color images at high quality by a holographic display. *Proc. SPIE Practical Hologr.* **XX,** 6136 (2006).
7. Jackel, J., Rice, C. & Veselka, J. Proton exchange for high-index waveguides in $LiNbO_3$. *Appl. Phys. Lett.* **41,** 607–608 (1982).
8. Matteo, A., Tsai, C. & Do, N. Collinear guided wave to leaky wave acoustooptic interactions in proton-exchanged $LiNbO_3$ waveguides. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **47,** 16–28 (2000).
9. Rust, U. & Strake, E. Acoustooptical coupling of guided to substrate modes in planar proton-exchanged $LiNbO_3$-waveguides. *Proc. Integrated Photonics Research* ME4 (Vol. 10 OSA Technical Digest Series, Optical Society of America, 1992).
10. Onural, L., Bozdagi, G. & Atalar, A. A new holographic 3-dimensional television display. *Proc. 1991 IEEE Ultrason. Symp.* **1,** 543–546 (1991).
11. Proklov, V. & Korablev, E. Multichannel waveguide devices using collinear acoustooptic interaction. *Proc. 1992 IEEE Ultrason. Symp.* **1,** 173–178 (1992).
12. Ito, K. & Kawamoto, K. An optical deflector using collinear acoustooptic coupling fabricated on proton-exchanged $LiNbO_3$. *Jpn. J. Appl. Phys.* **37,** 4858–4865 (1998).
13. Lee, H. The scophony television receiver. *Nature* **142,** 59–62 (1938).
14. Do, N. T., Su, J., Yoo, J., Matteo, A. M. & Tsai, C. S. High-efficiency acoustooptic guided-mode to leaky-mode conversion in proton-exchanged lithium niobate waveguides. *Proc. 1999 Ultrason. Symp.* 613–616 (1999).
15. Ohmachi, Y. & Noda, J. $LiNbO_3$ TE-TM mode converter using collinear acoustooptic interaction. *IEEE J. Quantum Electron.* **13,** 43–46 (1977).
16. Sohler, W. Integrated optics in $LiNbO_3$. *Thin Solid Films* **175,** 191–200 (1989).
17. Smalley, D. *et al.* Holovideo for everyone: a low-cost holovideo monitor. *J. Phys. Conf. Ser.* **415,** 012055 (2013).
18. Bove, V., Plesniak, W., Quentmeyer, T. & Barabas, J. Real-time holographic video images with commodity PC hardware. *Proc. SPIE Stereosc. Displays Applicat.* **XII,** 255262 (2005).
19. Barabas, J., Smithwick, Q., Smalley, D. & Bove, V. M. Real-time shader rendering of holographic stereograms. *Proc. SPIE Practical Hologr.* **XXIII,** 723303 (2009).
20. Smithwick, Q., Barabas, J., Smalley, D. & Bove, V. M. Interactive holographic stereograms with accommodation cues. *Proc. SPIE Practical Hologr.* **XXIV,** 761903 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by consortium funding at the MIT Media Lab and by Intel Corp. Graphics hardware was provided by NVIDIA. D.E.S. thanks the MIT Nanostructures Laboratory for use of its facilities, E. Pearson for discussion, and D. Novy for assistance with GPU coding.

Author Contributions D.E.S. performed experimental work and fabricated devices. D.E.S., Q.Y.J.S. and V.M.B. participated in conceptualization of waveguide phenomena for holographic video. D.E.S., Q.Y.J.S., V.M.B., J.B. and S.J. participated in the design and evaluation of experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.E.S. (desmalley@gmail.com).

METHODS

Proton-exchanged waveguide. The proton-exchange process is illustrated in Supplementary Fig. 1. An *x*-cut lithium niobate wafer 1 mm thick was cleaned using a standard cleaning process (3:1:1 ammonium hydroxide, hydrogen peroxide and water heated to 80 °C), rinsed in deionized water and then with a solvent such as isopropanol (IPA) to prevent residue formation during drying. Physically enhanced chemical vapour deposition (PECVD) was used to deposit a 200 nm silicon dioxide layer on the wafer. Negative resist (Futurrex NR8-1000) was spun on at 3,000 r.p.m. and the wafer was pre-baked in an oven at 100 °C for 7 min. The pattern was exposed with a light-field mask to define the waveguides, the resist was developed in 2% tetramethylammonium hydroxide (TMAH) solution and the underlying silicon dioxide etched in a buffered oxide etch for 30 s. Resist was removed with acetone. Benzoic acid was heated to 238 °C (it is recommended that the melt be diluted with 1% lithium benzoate by weight, but this was not done for the present work) and the wafer carefully placed in the melt for 34 min. (Experience shows that the wafer must be warmed gradually before entering the melt or it may break; lowering it to just above the melt surface allows it to warm to the melt temperature.) The wafer was removed carefully and slowly, to avoid cracking, cooled and cleaned with acetone and IPA. Silicon dioxide was removed by submerging the wafer in buffered oxide etch for 30 s, then the wafer was placed in a covered quartz dish and baked for 45 min in an oven preheated to 375 °C.

Al transducers. The lift-off process is illustrated in Supplementary Fig. 2. On a clean proton exchanged substrate, 600 nm of poly(methyl methacrylate) (PMMA) was spun, then the substrate was baked at 150 °C for 15 min. A layer of E-spacer (Showa Denko) or Aquasave (Mitsubishi Rayon) was then spun on to prevent charging while direct writing with an electron beam (we note that this could also be accomplished with a 20-nm-thick evaporated layer of chrome which would have to be stripped before development, but this was not done for the present work). An electron beam was used to direct write the transducer pattern at a dose of approximately 250 μC cm^{-2}. For the device used in this paper, the transducer was the composed of three regions each with a uniform period corresponding to 270 MHz, 310 MHz and 380 MHz. (These frequencies will vary with proton exchange time and temperature. Also, note that the features of these transducers are large enough to be patterned by photolithography if desired, but electron beam direct write allows for a high degree of customization and is convenient for small samples.) The Aquasave or E-spacer was removed from the exposed sample with deionized water, and the PMMA developed in a 1:1 mixture of IPA:MIBK (methyl isobutyl ketone) for approximately 30 s. A 200 nm film of aluminium was deposited by e-beam evaporation, and the sample placed in *N*-methyl-2-pyrrolidone (NMP) heated to 50 °C, and left until the Al lifted off. (Sonication at low power for 5 s may be required.) The exit face of the sample was polished down to a 0.3 μm grit, and the sample cleaned with acetone, methanol and isopropyl alcohol. The transducers were wire-bonded, using 2 thousandths of an inch thick aluminium wire, to a copper PCB board equipped with a 50 Ω radio-frequency connector. Impedance matching for the highest resonance was achieved with a lumped element L network (typically our samples required a 100 nH series inductor followed by a 9 pF shunt capacitor); matching the highest resonance was done to make up for the fact that the blue interaction is the least efficient.

Experiments. Polarization rotation. To demonstrate polarization rotation, light from a diode laser at $\lambda = 633$ nm was evanescently coupled into an anisotropic leaky mode device using a rutile prism. The output of the device was scanned with a *x*–*y* scanner onto a camera sensor (a camera with the lens removed) to allow for lower ISO images and less camera noise. A polarizer was placed at the output of device.

Frequency multiplexing of colour. The mode coupling frequency response for red, green and blue light was measured by coupling laser light into the TE$_1$ guided mode of the device and then exciting an acoustic wave with a radio-frequency signal which swept from 150 to 500 MHz. The light that was coupled into the leaky mode was measured with a light meter. This process was repeated for red, green and blue. Note that the shape of the device's frequency response represents not only the frequency response of the anisotropic interaction alone but also the response of the SAW transducer and the impedance matching network which was designed to give the best match at frequencies responsible for blue mode coupling. The power of the input light was 10 mW for red and 100 mW for green and blue. Holographic stereograms. The holographic stereogram images were created by taking one stereogram view at a resolution of 296 pixels× 156 pixels, stretching its resolution to 29,600 pixels × 156 pixels, and finally stitching 12 of these images together for a composite resolution of 355,200 pixels × 156 pixels. The α values of each of the red, green and blue channels of this image were multiplied by a different sinusoidal pattern in an OpenGL shader. All three colour signals were summed and divided by three, and sent out via one of the video card outputs (for example, the nominal 'red' channel). This signal was then up-converted and amplified before entering a single transducer of the modulator array.

For holographic stereogram images, light from three lasers (at $\lambda = 445$ nm, $\lambda = 532$ nm and $\lambda = 633$ nm) was combined in an X cube and focused with an achromatic lens into one channel of an anisotropic leaky mode coupling array. The output of the device was spatially filtered and focused on to the face of a spinning polygon (to optically descan the holographic fringe pattern so that it would appear stationary), vertically scanned onto a parabolic mirror, and finally imaged by a camera. For simplicity, only the view entering the camera was computed and displayed. The vertical diffuser, which extends the vertical view-zone of HPO holograms, was not used.

doi:10.1038/nature12296

The linear ubiquitin-specific deubiquitinase gumby regulates angiogenesis

Elena Rivkin[1,2]*, Stephanie M. Almeida[1,2]*, Derek F. Ceccarelli[1]*, Yu-Chi Juang[1]*, Teresa A. MacLean[1,2], Tharan Srikumar[3], Hao Huang[1], Wade H. Dunham[1,2], Ryutaro Fukumura[4], Gang Xie[1], Yoichi Gondo[4], Brian Raught[3], Anne-Claude Gingras[1,2], Frank Sicheri[1,2] & Sabine P. Cordes[1,2]

A complex interaction of signalling events, including the Wnt pathway, regulates sprouting of blood vessels from pre-existing vasculature during angiogenesis. Here we show that two distinct mutations in the (uro)chordate-specific gumby (also called *Fam105b*) gene cause an embryonic angiogenic phenotype in *gumby* mice. Gumby interacts with disheveled 2 (DVL2), is expressed in canonical Wnt-responsive endothelial cells and encodes an ovarian tumour domain class of deubiquitinase that specifically cleaves linear ubiquitin linkages. A crystal structure of gumby in complex with linear diubiquitin reveals how the identified mutations adversely affect substrate binding and catalytic function in line with the severity of their angiogenic phenotypes. Gumby interacts with HOIP (also called RNF31), a key component of the linear ubiquitin assembly complex, and decreases linear ubiquitination and activation of NF-κB-dependent transcription. This work provides support for the biological importance of linear (de)ubiquitination in angiogenesis, craniofacial and neural development and in modulating Wnt signalling.

During angiogenesis, new blood vessels sprout from pre-existing vasculature to form a microcapillary network and become uniquely adapted to the physiology and function of the organs that they infiltrate (reviewed in ref. 1). Several well-characterized molecular pathways direct vascular patterning, but contributions of Wnt pathways are just emerging. Wnt signalling pathways control a broad spectrum of events, including cell-fate specification, proliferation and migration (reviewed in ref. 2), and are grouped into the canonical, β-catenin-dependent and non-canonical pathways. Both canonical and non-canonical Wnt pathways influence angiogenesis (reviewed in refs 3, 4). Canonical Wnt signalling is required for angiogenesis in the CNS[5,6], whereas non-canonical Wnt signalling has been implicated in global early embryonic angiogenesis[7]. The events that occur in Wnt-responsive cells depend, in part, on dishevelled and its associated proteins to transmit stimuli to the appropriate pathway(s) (reviewed in ref. 8).

Protein modifications, such as ubiquitination, offer a rapid mechanism by which cells integrate inputs from signalling pathways. Ubiquitin is an evolutionarily conserved 76-amino-acid protein used to label and alter the fate of cellular proteins. During protein ubiquitination, a covalent bond is generated most commonly between the most carboxy-terminal amino acid residue, a glycine, and a lysine in the modified protein. Ubiquitin itself contains seven lysine residues that can undergo ubiquitination to form polyubiquitin oligomers (chains). These chains serve to direct protein localization, stability and activity. The linear ubiquitin assembly complex (LUBAC), comprised of HOIL-1/1l, HOIP (also called RNF31) and sharpin proteins, mediates formation of an atypical ubiquitin chain topology[9] involving the linear linkage of the C terminus of glycine 76 in one ubiquitin protomer to the free amino terminus of methionine 1 in a second protomer (reviewed in refs 10–12). So far, linear ubiquitination has been shown to regulate NF-κB-dependent inflammation and adaptive immunity[13–16]. Deubiquitination, the selective removal of ubiquitins, is equally important and is performed by members of five distinct deubiquitinase (DUB) families, one of which is the ovarian tumour (OTU) domain containing protease family[17]. Currently, no DUB has been shown to cleave linear ubiquitin chains exclusively.

Here, our studies of the recessive, embryonically lethal *gumby* mouse mutant and its angiogenic phenotype have led to the identification of the affected (uro)chordate-specific gene, *Fam105b* (which we call gumby (*Gum*)). We show that gumby encodes a linear ubiquitin-specific DUB that structurally belongs to the OTU family. Gumby can associate with LUBAC and counteract known LUBAC functions. We identify a role for the gumby–LUBAC axis in regulating canonical Wnt signalling. Our findings highlight the importance of linear (de)ubiquitination in angiogenesis, craniofacial and neuronal development, and Wnt signalling.

The *gumby* mutation causes embryonic angiogenic deficits

gumby/*gumby* mice were identified based on abnormal sprouting of the facial nerve at embryonic day (E)10.5[18], appear normal before E11.5, but die between E12.5–E14. Because shared molecular mechanisms can guide axons and blood-vessel branching, we examined vascular development in E10.0–11.0 +/+, *gumby*/+ and *gumby*/*gumby* embryos by whole-mount immunohistochemistry with platelet endothelial cell adhesion molecule-1 (PECAM-1) antibody (Supplementary Fig. 1). The major structures of the vascular system appeared similar in controls and mutants. However, branching vascular networks in the head and trunk were improperly organized and less complex in *gumby* homozygotes. In the medial region of the embryonic head, several large-diameter cranial vessels branch to form a hierarchical vascular network (Supplementary Fig. 1a, b). In *gumby*/*gumby* embryos, large cranial vessels were dilated, branching reduced and endothelial cells accumulated at branch points (Supplementary Fig. 1e, f and Fig. 1i). Normally, a 'capillary' network, the perineural vascular plexus (PNVP), forms in the trunk between intersomitic vessels and

[1] Samuel Lunenfeld Research Institute, Mt Sinai Hospital, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada. [2] Department of Molecular Genetics, University of Toronto, 1 King's Crescent, Toronto, Ontario M5S 1A8, Canada. [3] Department of Medical Biophysics, University of Toronto and Ontario Cancer Institute, Princess Margaret Cancer Centre, Toronto, Ontario M5G 1L7, Canada. [4] Mutagenesis and Genomics Team, RIKEN BioResource Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan.

*These authors contributed equally to this work.

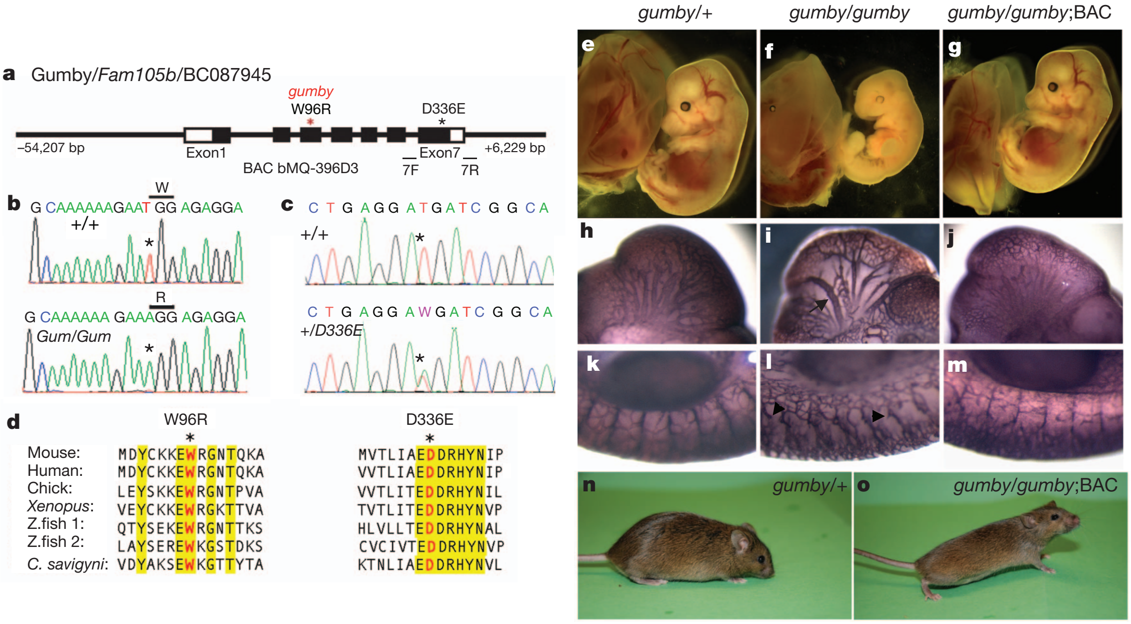

a Gumby/*Fam105b*/BC087945
gumby
W96R
D336E
–54,207 bp
Exon1
BAC bMQ-396D3
Exon7
7F
7R
+6,229 bp
b G CAAAAAAGAAT GG AGAGGA
W
+/+
R
G CAAAAAA GAAAGG AGAGGA
Gum/Gum
c C T G A G G A T G A T C G G C A
+/+
C T G A G G A W G A T C G G C A
+/*D336E*
d W96R
D336E
Mouse: MDYCKKEWRGNTQKA MVTLIAEDDRHYNIP
Human: MDYCKKEWRGNTQKA VVTLIAEDDRHYNIP
Chick: LEYSKKEWRGNTPVA VVTLITEDDRHYNIL
Xenopus: VEYCKKEWRGKTTVA TVTLITEDDRHYNVP
Z.fish 1: QTYSEKEWRGNTTKS HLVLLTEDDRHYNAL
Z.fish 2: LAYSEREWKGSTDKS CVCIVTEDDRHYNVP
C. savigyni: VDYAKSEWKGTTYTA KTNLIAEDDRHYNVL
gumby/+
gumby/gumby
gumby/gumby;BAC
e
f
g
h
i
j
k
l
m
n *gumby/+*
o *gumby/gumby*;BAC

Figure 1 | Identification of the *gumby* (*Gum*^{W96R}) causative mutation and the new *Gum*^{D336E} allele. a, Schematic diagram of the gumby gene and BAC bMQ-396D3. Primers 7F and 7R used to identify mice carrying *Gum*^{D336E} flank exon 7. **b**, **c**, Sequencing traces from +/+ and *gumby/gumby* (*Gum/Gum*) (**b**) and +/+ and *Gum*^{D336E}/+ (+/D336E) (**c**) mice. **d**, Amino acid sequence alignment spanning Gum^{W96R} and Gum^{D336E} mutations. Trp 96 and Asp 336 are shown in red, marked with asterisks. Yellow indicates interspecies amino acid sequence identity. **e**–**o**, BAC rescue of lethality and vascular abnormalities of *gumby/gumby* mice. Morphological appearance of *gumby*/+ (**e**), *gumby/gumby* (**f**) and *gumby/gumby* embryos carrying the BAC transgene (*gumby/gumby*;BAC) (**g**) at E14.0. **h**–**m**, PECAM-1 immunohistochemistry of *gumby*/+ (**h**, **k**), *gumby/gumby* (**i**, **l**) and *gumby/gumby*;BAC (**j**, **m**) embryos at E10.5. Cranial vasculature (**h**–**j**). Trunk vasculature (**k**–**m**) is shown at the forelimb bud level. Arrow indicates dilated cranial vessels; arrowheads highlight stunted branches in the trunk. **n**, **o**, Adult *gumby*/+ (**n**) and *gumby/gumby*;BAC (**o**) littermates.

extends into the neurectoderm (Supplementary Fig. 1c, d)[19]. In *gumby/gumby* embryos, fewer and less elaborate vessel extensions formed between the somites and the PNVP (Supplementary Fig. 1g, h and Fig. 1l).

Gum^{W96R} causes the *gumby* phenotypes

Meiotic mapping of 154 progeny from *gumby*/+ intercrosses refined the critical interval to 1.7 megabases (Mb) between D15Mit18 (26.7 Mb) and single nucleotide polymorphism (SNP) rs13482490 (28.4 Mb) on mouse chromosome 15 (Ensembl assembly v35) (Supplementary Fig. 2). Sequencing genes within this interval identified a T-to-A transversion (T285A) in the third exon of the *Fam105b* (also called BC087945) gene, which substitutes tryptophan at position 96 to arginine, and is referred to as *Gum*^{W96R} (Fig. 1b). Tryptophan 96 is conserved in all known orthologues (Fig. 1d). Whereas *Ciona intestinalis* and *Ciona savigyni* genomes each carry a copy, related genes are absent in non-chordates. We hereafter refer to *Fam105b* as the gumby (*Gum*) gene.

To test whether this is the *gumby* causative gene, we performed rescue experiments with bacterial artificial chromosome (BAC) bMQ-396D that spans the gumby gene and ~60 kilobases (kb) of its flanking region (Fig. 1a, e–o and Supplementary Fig. 3). One founder BAC transgenic line rescued the lethality (Fig. 1 g, o) and vascular deficits (Fig. 1j, m) of *gumby/gumby* mice. Thus, this mutation causes the *gumby* phenotype, and we refer to this gumby allele as *Gum*^{W96R}.

Identification and analysis of gumby allele *Gum*^{D336E}

We identified mice carrying another allele of gumby, *Gum*^{D336E}, by screening a bank of cryopreserved sperm and genomic DNA from over 6,000 G1 male progeny from *N*-ethyl-*N*-nitrosourea mutagenized G0 males (http://www.brc.riken.jp/lab/gsc/mouse/index.html)[20]. In the *Gum*^{D336E} allele, a T-to-A transversion in exon 7 changes conserved aspartate 336 to glutamate (Fig. 1c). Both *Gum*^{W96R} and *Gum*^{D336E} homozygotes show reduced branchial arches and embryonic lethality after E12.5 (data not shown). Using anti-PECAM-1 whole-mount immunofluorescence, we quantified the relative deficits in the cranial vasculature of *Gum*^{D336E} and *Gum*^{W96R} homozygotes at E10.5 (Fig. 2a–f). We found decreased numbers of secondary and tertiary vessels branching off the internal carotid artery (ICA) in *Gum*^{W96R} (Fig. 2c, g) and *Gum*^{D336E} (Fig. 2e, g) homozygotes relative to +/+ littermates (Fig. 2a, g). We examined vessel dilation by measuring the diameter of the ICA before its migration to the posterior head region (Fig. 2b, d, f, h) and secondary branch dilation by measuring the diameter of the first branch off the ICA (Fig. 2b, d, f, i). *Gum*^{W96R} homozygotes had larger dilated ICAs (Fig. 2h) and secondary branches (Fig. 2i) compared to +/+ and *Gum*^{D336E} homozygotes. Immunoblot and immunofluorescence experiments indicate that the *Gum*^{W96R} and *Gum*^{D336E} mutations do not detectably compromise gumby protein level or cytoplasmic localization (Supplementary Fig. 4). These findings further support an angiogenic requirement for gumby and predict that the *Gum*^{W96R} mutation affects protein function more severely than *Gum*^{D336E}.

Gumby messenger RNA is transcribed in the developing vasculature and other regions affected in mutants, including branchial arches and neural crest cells (Fig. 2j–l and Supplementary Fig. 5). The walls of blood vessels are composed of endothelial cells, which line the luminal vessel surface, and perivascular cells, which encircle the outside of the vascular endothelium[21]. In immunofluorescence experiments on E10.5–E11.5 embryonic sections with an anti-gumby antibody (Supplementary Fig. 6), gumby protein co-localized with the endothelial-specific marker PECAM-1, but not perivascular markers smooth muscle actin and desmin (Fig. 2m–p and data not shown). Gumby protein is enriched in a subset of endothelial cells. Gumby-enriched endothelial cells were present throughout intersomitic vessels (ISVs) and, in the PVNP and cranial vasculature, often located near possible leading regions of vessels (arrows, Fig. 2n, p) or at 'vascular buds' (arrowheads, Fig. 2n, p).

Gumby is a DUB that cleaves linear ubiquitin chains

Gumby (FAM105B) was identified in the NCBI database as a putative member of the OTU-fold DUBs, a prediction that we confirmed by

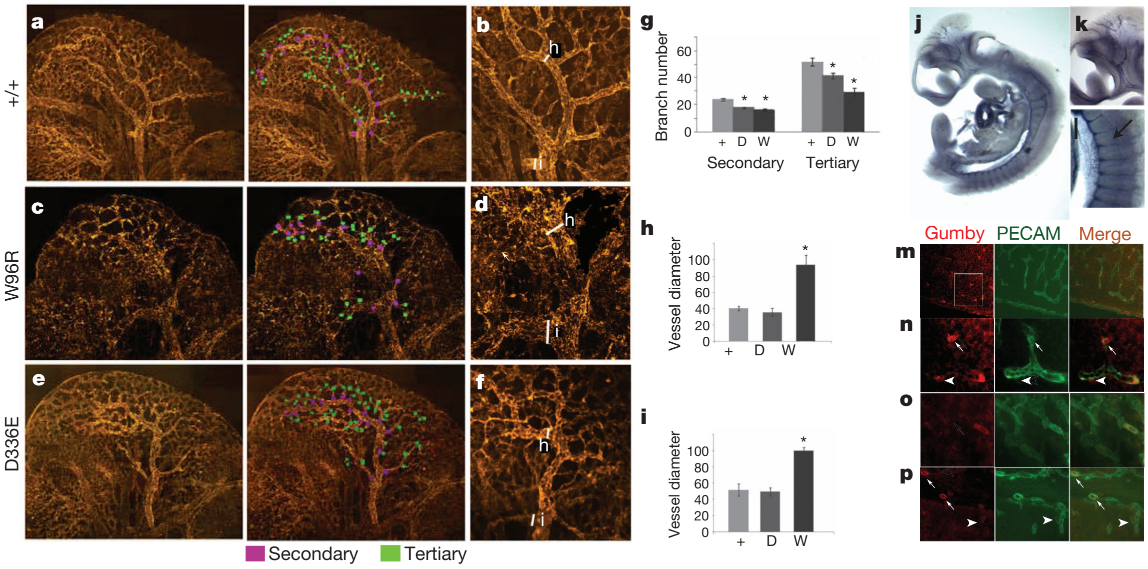


Figure 2 | Analyses of vascular phenotypes and gumby expression. **a**–**f**, Whole-mount anti-PECAM-1 immunofluorescence showing cranial vasculature of E10.5 +/+ (**a**, **b**), *Gum*^{W96R}/*Gum*^{W96R} (W96R) (**c**, **d**) and *Gum*^{D336E}/*Gum*^{D336E} (D336E) (**e**, **f**) embryos. **a**, **c**, **e**, Left: composite whole-head images; right: secondary and tertiary vessels are marked in purple and green, respectively. **b**, **d**, **f**, Analyses of cranial vessel dilation of intercranial artery (ICA). White bars indicate the areas of the ICA (**h**) and first branch (**i**) measured. **g**, Quantification of secondary and tertiary vessels branching off the primary ICA per side of head. **h**, **i**, Quantification of ICA (**h**) and first ICA branch (**i**) diameters. Asterisk indicates statistical significance relative to +/+ embryos ($n = 3$ or 6, $P < 0.05$, Student's *t*-test). **j**–**l**, RNA *in situ* hybridization of E10.5 +/+ embryos detects gumby RNA in the vasculature. Magnified views from head (**k**) and trunk (**l**) are shown. An intersomitic vessel (ISV) is marked with an arrow. **m**–**p**, Immunofluorescence with anti-PECAM-1 (green) and anti-gumby (red) antibodies detects gumby in endothelial cells in cross-sections of the trunk in +/+ E11.5 embryos. **n**, High magnification of boxed area in **m**. Gumby protein is enriched near presumptive tips of vessels (arrows) and vascular buds (arrowheads).

X-ray crystallographic analysis to 1.6 Å resolution (Supplementary Table 1). With 14.9% sequence identity (Supplementary Fig. 7) and a superimposable catalytic triad (Cys 129, His 339 and Asn 341), gumby most closely resembled the catalytic domain of OTUB1 (root mean squared deviation (r.m.s.d.) = 2.09 Å)[22–24], a DUB that specifically cleaves K48-linked ubiquitin chains[22,25] (Fig. 3a). As such, we examined DUB activity against all eight diubiquitin chain linkages *in vitro* (Fig. 3b, right panel). In contrast to OTUB1, gumby displayed selectivity for linear diubiquitin (Fig. 3b left panel). DUB activity was dependent on Cys 129, was not measurable against ubiquitin–AMC, and was insensitive to inhibition by ubiquitin–vinyl sulphone (Supplementary Fig. 8a–c). Unlike OTUB1 (refs 22, 25), residues preceding the catalytic domain were not required for DUB activity (compare Gum^{M55} and Gum^{R79} in Supplementary Fig. 8d). The gumby mutations Trp96Arg and Asp336Glu mapped to a surface of the OTU-fold 14.1 Å and 11.4 Å, respectively, from the catalytic cysteine (Fig. 3a), positions expected to have an impact on catalytic function. In confirmation, kinetic analysis of the W96R mutant revealed ~10,000- and 5-fold perturbations in k_{cat} and K_m, respectively, whereas the D336E mutant revealed ~50-fold perturbation in k_{cat} (Supplementary Fig. 8d). These effects on enzymatic function paralleled the relative severity of mutant phenotypes *in vivo* (Fig. 2a-i).

We solved the 2.4 Å and 2.8 Å crystal structures of gumby in the presence of free ubiquitin and linear diubiquitin substrate, respectively (Supplementary Table 1). Linear diubiquitin engaged an extensive surface (total buried surface area = 3,267 Å^2) on gumby with the active site Cys 129 centrally positioned at the scissile bond linking Gly 76 and Met 1 of distal and proximal ubiquitin moieties (Fig. 3a). Monoubiquitin engaged the distal ubiquitin-binding site of gumby similar to the mode observed in the diubiquitin (Supplementary Fig. 9a) co-structure, but different from modes observed in yeast Otu1, viral OTU and human OTUB5 co-structures, possibly reflecting differences in substrate specificities (Supplementary Fig. 9b).

In the diubiquitin complex, distally bound ubiquitin engaged gumby through a combination of hydrophobic and polar interactions burying a large total surface area of 1,771 Å^2 (Fig. 3c, d). The contact surface on distal ubiquitin was composed of strands β3, β4, loops connecting β1–β2, β3–β4, and the C-terminal tail, whereas the reciprocal surface on gumby was composed of helix α8, loops between β2–α3, α9–α10 and β3–β4 (Fig. 3c and Supplementary Fig. 7). Proximally bound ubiquitin engaged gumby primarily through hydrophilic interaction, burying a total surface area of 1,496 Å^2. The interaction surface on proximally bound ubiquitin was composed of strands β1, β2, helix α1 and loops between β2–α1 and β3–β4, whereas the reciprocal contact surface on gumby was composed of helices α1, α2, loop regions β2–α3, α9–α10 and β5–β6. A detailed list of contact residues and side-chain interactions between linear diubiquitin and gumby are shown in Fig. 3d and Supplementary Fig. 9c, d.

The prominent position of Trp 96 and Asp 336 in the general vicinity of the active site and on the direct contact surface with diubiquitin substrate allowed rationalization of the differential effects of the Trp96Arg and Asp336Glu gumby mutations. As a generally conserved hydrophobic position in OTU sequences, Trp 96 contributes to enzyme structure in addition to mediating a direct contact with Lys 33 of the proximally bound ubiquitin (Fig. 3d and Supplementary Fig. 9d). Mutation of the partially buried Trp side chain to Arg would perturb both local structure and substrate binding affinity, manifesting in the observed changes to both k_{cat} and K_m. Consistent with this inference, mutation of the topologically equivalent hydrophobic position in OTUB1 (Tyr61Arg) abolished all trace of catalytic function (Supplementary Fig. 8e), and direct binding experiments showed a ~20-fold loss in substrate affinity for the Trp96Arg gumby mutant (Supplementary Fig. 8f). From its highly solvent-exposed position in the *apo* structure, Asp 336 contributes less to enzyme structure than Trp 96 while forming a direct salt interaction with Lys 29 of the proximal site ubiquitin. Thus, the conservative Asp336Glu mutation would impact less on active site structure than the Trp96Arg mutation while preserving a favourable salt interaction with Lys 29 on ubiquitin, manifesting in the observed smaller change to k_{cat} and no change to K_m.

Comparison of gumby with OTUB1 (Protein Data Bank accession 4DDG)[23] allows for a rationalization of its unique substrate specificity and regulation (Supplementary Fig. 9e). Despite using equivalent

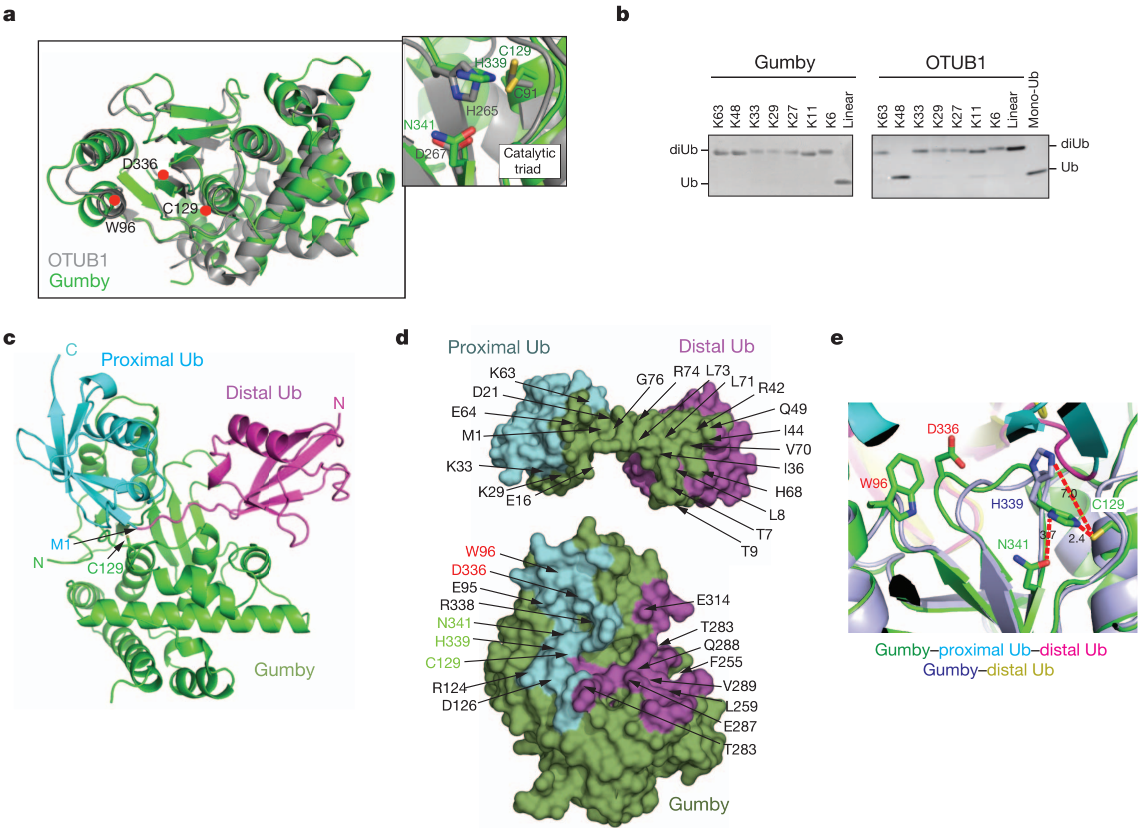


Figure 3 | Structural and biochemical analysis of gumby. **a**, Ribbons view of *apo*-gumby (green) superimposed on OTUB1 (Protein Data Bank 4DDG) (grey). Inset shows catalytic triad. **b**, Cleavage specificity towards diubiquitin chains. Ub, ubiquitin. **c**, Ribbons view of the gumby–linear diubiquitin complex. **d**, Peel-away surface views of linear diubiquitin (top) bound to gumby (bottom). Domains coloured as in **c**, with residues involved in intermolecular contacts coloured according to the domains contacted. Catalytic triad and mutant positions are labelled green and red, respectively. **e**, Active-site comparison of gumby–diubiquitin and gumby–monoubiquitin complexes. A productive orientation of His 339 is observed only in the gumby–diubiquitin complex.

topological surfaces on their OTU folds for substrate engagement (centre of mass positions for distal and proximal ubiquitins differ by only 1.6 and 3.9 Å, whereas angles of rotation differ more greatly by 60° and 135°, respectively), only 2 of 45 total contact residues on gumby, namely Cys 129 and His 339, were conserved in OTUB1. Because both gumby and OTUB1 engage continuous and extensive surfaces spanning both ubiquitin molecules in their respective diubiquitin substrates, centred on the site of cleavage, this would greatly constrain the ability of each enzyme to recognize substrates with alternative linkage topologies, which present vastly different interaction surfaces. Similar to OTUB1, a productive orientation of the catalytic triad is only apparent for gumby bound to diubiquitin substrate. The coupling of catalytic activation with the binding of preferred substrate (Fig. 3e) would further accentuate specificity by prohibiting cleavage of partially and/or suboptimally engaged substrates.

Gumby interacts with and can counteract LUBAC

We identified the LUBAC component HOIP as a gumby interactor by affinity purification of Flag-tagged gumby and its associated proteins followed by mass spectrometry (AP–MS). Whereas full-length gumby possesses a PDZ binding motif (PBM) and interacts with multiple PDZ-domain-containing proteins when expressed in HEK293 cells (data not shown), a gumby construct containing a mutated PBM consisting of four sequential alanine residues (gumby$^{\Delta PBM}$) (Fig. 4a) interacted with HOIP (Fig. 4b and Supplementary Table 2). No other LUBAC components were detected, suggesting that the interaction with HOIP may be independent of other delineated LUBAC components. By performing immunoprecipitation coupled to immunoblotting using Flag–gumby expression constructs, we confirmed this interaction and mapped it to the N terminus of gumby. A construct truncated at position 105 (gumby^{C105X}) sufficed to mediate the interaction, whereas one missing the first N-terminal 54 residues (gumby$^{\Delta 54}$) could not interact with HOIP. HOIP interacted with wild-type gumby and gumby^{C129S} proteins equivalently (Fig. 4c–e). Thus, gumby interacts through its N terminus with HOIP independently of its PBM or catalytic DUB domain.

Next we tested whether gumby could affect linear ubiquitination of proteins *in vivo*. Co-expressing HOIL and HOIP in HEK293T cells increased the level of linearly ubiquitinated proteins in immunoblot and immunofluorescence experiments using an anti-linear ubiquitin antibody[26] (Fig. 4f–h and Supplementary Fig. 10). Co-transfection of Flag–gumby—but not Flag–gumby^{C129S} and Flag–gumby^{W96R}—with HOIL and HOIP decreased overall levels of proteins modified with linear ubiquitin (Fig. 4f and Supplementary Fig. 10). In mice, levels of linearly ubiquitinated proteins were elevated in E10.0 *Gum*^{W96R}/*Gum*^{W96R} embryos relative to their +/+ and heterozygous littermates, as detected by immunoblot analyses using an anti-linear ubiquitin antibody (Fig. 4i). Thus, the *Gum*^{W96R} mutation compromises linear deubiquitination in cell culture and animals.

Finally, we examined whether gumby could oppose LUBAC-dependent activation of NF-κB-directed transcription (Fig. 4j, k). Co-transfecting an NF-κB-dependent reporter construct with gumby or gumby$^{\Delta PBM}$ caused ~100-fold inhibition of LUBAC-dependent activation in HEK293T cells. Gumby^{C129S}, gumby^{W96R} and gumby^{C105X} could not repress this activation to an equivalent level. Neither LUBAC components nor gumby in combination or alone affects AP2-dependent transcription (data not shown), suggesting that these are not effects on general transcription. Whereas gumby interacts with HOIP via its

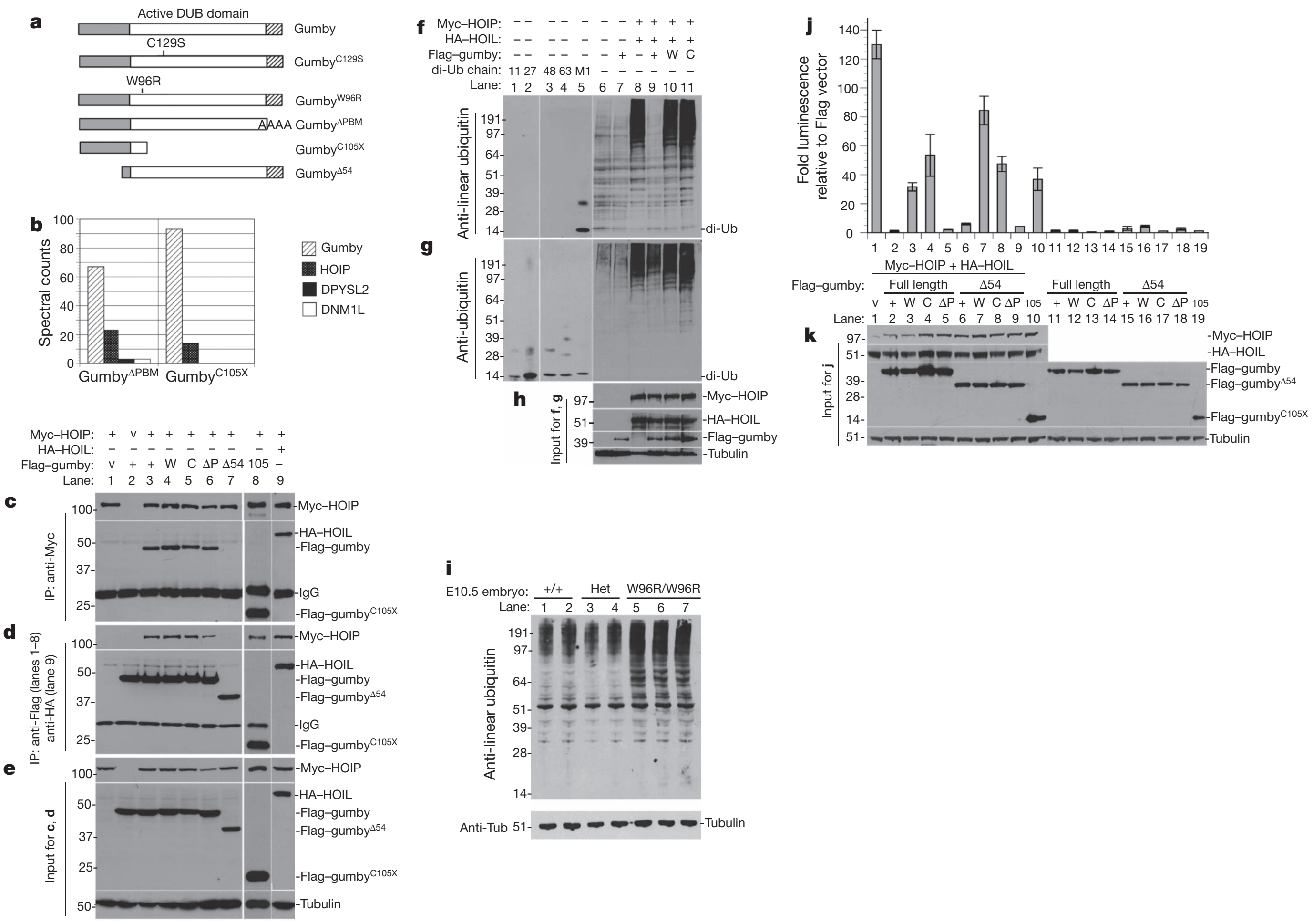


Figure 4 | Gumby interacts with HOIP and counteracts LUBAC activity
a, Flag-tagged gumby constructs used. **b**, Representative mass spectrometry results for Flag-tagged gumby$^{\Delta PBM}$ and gumby^{C105X} immunoprecipitations. HOIP was identified as the most significant interacting protein in HEK293 cells expressing Flag-tagged gumby$^{\Delta PBM}$ or gumby^{C105X}, for details see Methods and Supplementary Table 2. **c**–**e**, Immunoprecipitation coupled to immunoblotting using Flag–gumby expression constructs, HA–HOIL and Myc–HOIP determined that HOIP interacts with wild-type gumby (+), gumby^{W96R} (W), gumby^{C129S} (C), gumby$^{\Delta PBM}$ (P) and gumby^{C105X} (105) proteins, but not gumby$^{\Delta 54}$ (54). Flag–vector is designated as 'v'. **c**, Immunoblot of immunoprecipitation using anti-Myc antibody for Myc–HOIP. **d**, Immunoblot of immunoprecipitation using anti-Flag antibody for Flag-tagged gumby constructs. **e**, Inputs for immunoprecipitation experiments. **f**–**h**, Immunoblot using anti-linear ubiquitin antibody detected decreased levels of linear ubiquitinated proteins in HEK293T cells, when gumby, but not gumby^{C129S} or gumby^{W96R}, was co-expressed with HOIL and HOIP. The anti-linear ubiquitin antibody recognizes linear diubiquitin (M1), but not control diubiquitin chains K11 (11), K48 (48) or K63 (63). **g**, Equivalent levels of overall ubiquitination were detected in HOIL–HOIP co-expressing cells by anti-pan-ubiquitin antibody. **h**, Inputs for **f** and **g**. **i**, Immunoblot with anti-linear ubiquitin antibody detects increased levels of linear ubiquitinated protein in individual E10.5 *Gum*^{W96R}/*Gum*^{W96R} embryos (W96R/W96R) relative to +/+ and *Gum*^{W96R} /+ (Het) littermates. **j**, Catalytically active gumby counteracts HOIP–HOIL-dependent stimulation of luciferase expression from an NF-κB reporter in HEK293T cells. Wild-type, gumby$^{\Delta 54}$, gumby$^{\Delta PBM}$ or gumby$^{\Delta 54 \Delta PBM}$ fully suppresses this stimulation. Data are presented as means ± s.e.m. ($n = 3$; $P < 0.05$, Student's *t*-test). **k**, Immunoblot of inputs for luciferase assays. Units of markers shown along the left side of blots in **c**–**e**, **f**–**i** and **k** are in kilodaltons (kDa).

N-terminal region, in these assays, gumby$^{\Delta 54}$ can still antagonize LUBAC-dependent activation. Thus, in an overexpression system, the ability of gumby to fully offset LUBAC function requires its catalytic activity.

Gumby and LUBAC modulate Wnt signalling

Recovery of gumby as a dishevelled 2 (DVL2) interactor in a yeast two-hybrid screen suggested a role for gumby in Wnt signalling[27]. We confirmed the gumby–DVL2 interaction in HEK293T cells by immunoprecipitation of haemagglutinin (HA)-tagged DVL2, which recovered Flag-tagged gumby and gumby^{C129S} equivalently and reciprocally (Fig. 5a–c). Gumby$^{\Delta 54}$ could not interact with DVL2, whereas gumby^{C105X} could. Thus, gumby interacts with DVL2 via its N-terminal region.

Canonical Wnt signalling activates T-cell-specific transcription factor/lymphoid enhancer factor (TCF/LEF) to induce downstream target gene expression[2]. We tested whether gumby might modulate canonical Wnt3a signalling by assaying TCF/LEF-dependent luciferase expression of the TOPFLASH reporter in HEK293T cells (Fig. 5d–g). Flag–gumby could enhance luciferase expression in the presence of WNT3A. Flag–gumby^{W96R} or Flag–gumby^{C129S} enhanced WNT3A-induced TOPFLASH expression to a lesser degree. When we asked whether gumby and LUBAC might have opposing roles in Wnt signalling, we found that co-expressing HA–HOIL and Myc–HOIP together inhibited TOPFLASH expression over tenfold. Addition of Flag–gumby or gumby$^{\Delta 54}$ reversed HOIL–HOIP-dependent inhibition, but addition of Flag–gumby^{C129S} or gumby^{W96R} did not. The same trends were seen in the presence of DVL2 (Fig. 5 f, g). These data indicate that linear (de)ubiquitination via the gumby–LUBAC axis can modulate canonical Wnt signalling.

To determine whether gumby compromised Wnt response in endothelial cells, we took advantage of canonical Wnt pathway reporter mice (*TOPGAL*) that carry a transgene with LEF/TCF binding sites directing β-galactosidase (β-gal) expression[28]. In triple immunofluorescence experiments using antibodies against β-gal, gumby and PECAM-1 at signal saturation for β-gal, gumby protein is expressed exclusively in endothelial cells expressing β-gal in E10.5 +/+;*TOPGAL/+*, *Gum*^{W96R}/*Gum*^{W96R};*TOPGAL/+* and *Gum*^{D336E}/*Gum*^{D336E};*TOPGAL/+* embryos (Supplementary Fig. 11). Gumby/β-gal co-localization was also seen in the dorsal root ganglia and telencephalon (Fig. 5h, k, Supplementary Fig. 11, and data not shown). To test whether canonical Wnt signalling read-out might be affected in mutant endothelial cells, we analysed relative β-gal levels in ISVs, because, in these, all endothelial cells co-express β-gal and gumby (Supplementary Fig. 11) and their stereotypic organization facilitates comparative analyses. Under conditions where we detect intermediate β-gal signal in PECAM-1-labelled ISVs of E10.5 +/+;*TOPGAL/+* embryos, β-gal immunofluorescence was significantly reduced in *Gum*^{W96R}/*Gum*^{W96R};*TOPGAL/+* and *Gum*^{D336E}/*Gum*^{D336E};*TOPGAL/+* embryos (Fig. 5 h–o). Levels were also reduced in other gumby-expressing regions, including the dorsal root ganglia and telencephalon (data not shown). We note that *Gum*^{W96R} and *Gum*^{D336E} homozygotes show a CNS-specific angiogenic phenotype similar to that reported upon disruption of canonical Wnt signalling[5,6]. Thus, these findings suggest a role for gumby in canonical Wnt signalling during angiogenesis.

Discussion

Here we have shown that loss of the linear DUB activity of gumby compromises angiogenesis and neuronal and craniofacial development. So far, biological roles identified for LUBAC components in the mouse have been milder. Mutations in sharpin lead to chronic proliferative dermatitis[29,30]. *HOIL-1/1l* mouse mutants show no overt pathology unless challenged[31]. *HOIP* mutant phenotypes remain to be reported. Explanations for these observations include the existence of other LUBAC complexes or components, or unappreciated functional redundancy between sharpin and HOIL. Also, in the absence of gumby, constitutive linear ubiquitination of proteins could be more damaging than absence of linear ubiquitination. For instance, in cases where linear ubiquitination is inhibiting Wnt signalling, Wnt signalling could not be (re)activated by deubiquitination.

Our work reveals that gumby, via its N terminus, interacts with HOIP and DVL2. In overexpression experiments, the ability of gumby to fully counteract LUBAC function requires its catalytic activity. However, *in vivo*, at endogenous protein levels, the N-terminal interaction may have important functions—perhaps to regulate activity of the overall complex, guide substrate selection, direct subcellular localization or direct gumby to modulate specific pathways.

In humans, deletions of the gumby-containing interval on chromosome 5p15.2 are associated with mental retardation and craniofacial anomalies observed in patients with cri du chat syndrome (CdCS)[32–34]. Although not the only gene deleted, gumby might contribute to some CdCS symptoms via its effects on linear deubiquitination.

In summary, our work has identified gumby as a linear ubiquitin-specific DUB with roles in angiogenesis, other developmental processes and Wnt signalling. This paves the way for future investigations into the identification of gumby substrates, understanding of the roles

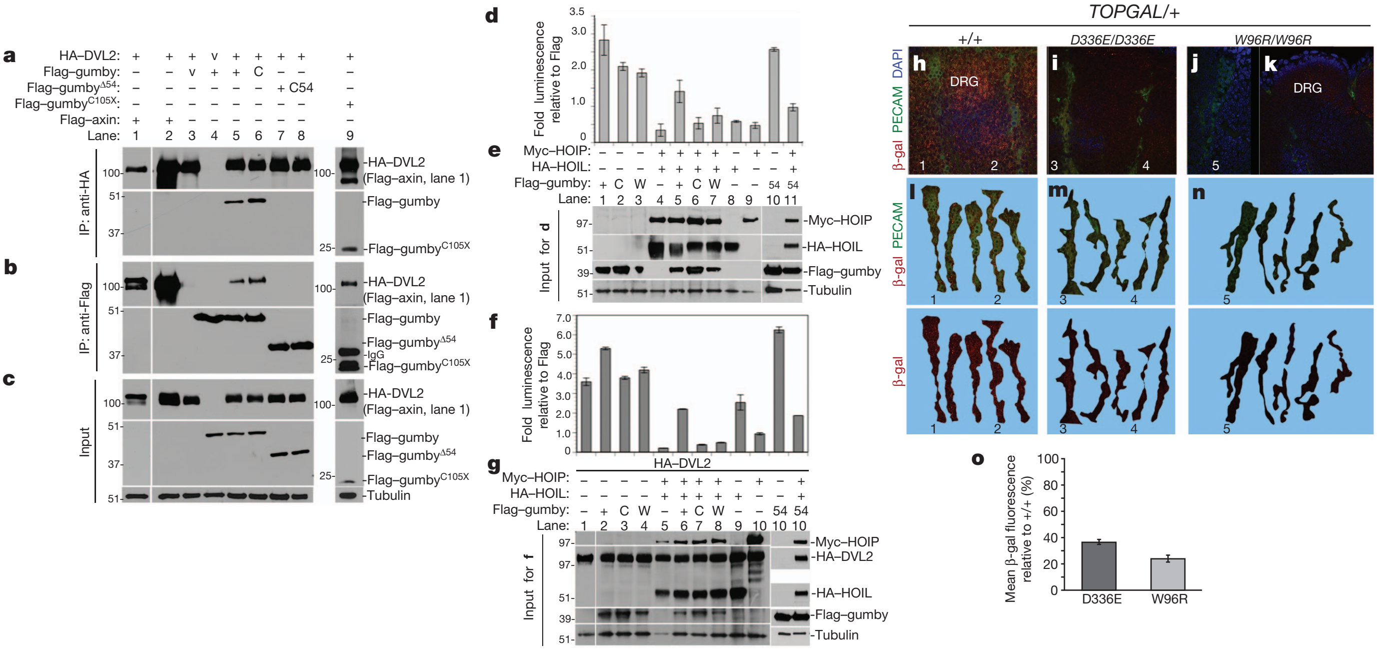


Figure 5 | Gumby interacts with DVL2 and can modulate Wnt signalling. **a**, Immunoprecipitation (IP) with anti-HA antibody of HA-tagged DVL2 (HA–DVL2) recovered Flag-tagged gumby, gumby^{C129S} (C), gumby^{C105X} and Flag–axin (positive control), but not gumby$^{\Delta54}$ (54) or gumby$^{\Delta54\ C129S}$ (C54). Flag–vector is designated as 'v'. **b**, In immunoprecipitation with anti-Flag antibody, all Flag–gumby constructs, except for Flag–gumby$^{\Delta54}$ or gumby$^{\Delta54\ C129S}$, recovered HA–DVL2. **c**, Input amounts and tubulin levels. **d**, **f**, In assays for WNT3A activation of luciferase expression from the TOPFLASH reporter performed in the absence (**d**, **e**,) or presence (**f**, **g**) of HA–DVL2, Flag–gumby enhanced TOPFLASH expression more than Flag–gumby^{W96R} (W) and Flag–gumby^{C129S}. HA–HOIL and Myc–HOIP co-expression inhibited TOPFLASH. Gumby, but not gumby^{C129S} or gumby^{W96R}, reversed HOIL–HOIP-dependent inhibition. **e**, **g**, Immunoblots show inputs for luciferase assays in **d** and **f**, respectively. Data are presented as means ± s.e.m., $n = 3$ or more ($P < 0.05$, Student's *t*-test). **h**–**o**, Analysis of β-galactosidase (β-gal) expression in ISVs of E10.5 embryos. **h**, Co-immunofluorescence with anti-β-gal (red) and anti-PECAM-1 (green) antibodies detects high β-gal within ISVs and dorsal root ganglia (DRG) of +/+;*TOPGAL/+* embryos. In *Gum*^{D336E}/*Gum*^{D336E};*TOPGAL/+* (**i**) and *Gum*^{W96R}/*Gum*^{W96R};*TOPGAL/+* (**j**, **k**) embryos, β-gal signal is lower in ISVs and in DRG. **l**–**n**, Representative traces of ISVs scored for β-gal fluorescence intensity are shown for +/+;*TOPGAL/+* (**l**) *Gum*^{D336E}/*Gum*^{D336E};*TOPGAL/+* (**m**) and *Gum*^{W96R}/*Gum*^{W96R};*TOPGAL+* (**n**) embryos. Numbers in **l**–**n** mark specific ISVs traced in **h**, **i** and **j**. Top panels show β-gal and PECAM-1 immunofluorescence. Bottom panels show only β-gal immunofluorescence. **o**, Graph of percentage mean fluorescence in ISVs of *Gum*^{D336E}/*Gum*^{D336E};*TOPGAL/+* (D336E) and *Gum*^{W96R}/*Gum*^{W96R};*TOPGAL+* (W96R) embryos (36.6 ± 1.8% and 24.0 ± 2.8%, respectively). Data are presented as means ± s.e.m. Effect of gumby genotype on anti-β-gal mean fluorescence intensity was highly significant (Kruskal–Wallis test ($k = 3$); $H = 43.69$; $P < 0.0001$). Units of markers shown along the left side of blots in **a**–**c**, **e** and **g** are in kilodaltons (kDa).

and regulation of the linear deubiquitination–ubiquitination balance and possibly development of antiangiogenic therapies.

METHODS SUMMARY

All mouse husbandry and handling was performed in conformity with the Canadian Council of Animal Care recommendations (AUP 0024a-00H). For all experiments, unless otherwise stated, a minimum of six homozygous mutants of each genotype and six +/+ littermates were examined. X-ray structure analyses, binding studies using ITC, deubiquitination assays, RNA and protein expression analyses, immunoprecipitations and Wnt reporter expression analyses *in vivo* and in cell culture were performed by standard methods and are described in detail in Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 3 November 2012; accepted 20 May 2013.
Published online 24 May 2013.

1. Risau, W. Mechanisms of angiogenesis. *Nature* **386,** 671–674 (1997).
2. Logan, C. Y. & Nusse, R. The Wnt signaling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.* **20,** 781–810 (2004).
3. Liebner, S. & Plate, K. H. Differentiation of the brain vasculature: the answer came blowing by the Wnt. *J. Angio. Res.* **2,** 1 (2010).
4. Zerlin, M., Julius, M. A. & Kitajewski, J. Wnt/Frizzled signaling in angiogenesis. *Angiogenesis* **11,** 63–69 (2008).
5. Daneman, R. *et al.* Wnt/β-catenin signaling is required for CNS, but not non-CNS, angiogenesis. *Proc. Natl Acad. Sci. USA* **106,** 641–646 (2009).
6. Stenman, J. M. *et al.* Canonical Wnt signaling regulates organ-specific assembly and differentiation of CNS vasculature. *Science* **322,** 1247–1250 (2008).
7. Cirone, P. *et al.* A role for planar cell polarity signaling in angiogenesis. *Angiogenesis* **11,** 347–360 (2008).
8. Wharton, K. A. Jr. Runnin' with the Dvl: proteins that associate with Dsh/Dvl and their significance to Wnt signal transduction. *Dev. Biol.* **253,** 1–17 (2003).
9. Kirisako, T. *et al.* A ubiquitin ligase complex assembles linear polyubiquitin chains. *EMBO J.* **25,** 4877–4887 (2006).
10. Behrends, C. & Harper, J. W. Constructing and decoding unconventional ubiquitin chains. *Nature Struct. Mol. Biol.* **18,** 520–528 (2011).
11. Walczak, H., Iwai, K. & Dikic, I. Generation and physiological roles of linear ubiquitin chains. *BMC Biol.* **10,** 23 (2012).
12. Iwai, K. Linear polyubiquitin chains: a new modifier involved in NFκB activation and chronic inflammation, including dermatitis. *Cell Cycle* **10,** 3095–3104 (2011).
13. Gerlach, B. *et al.* Linear ubiquitination prevents inflammation and regulates immune signalling. *Nature* **471,** 591–596 (2011).
14. Ikeda, F. *et al.* SHARPIN forms a linear ubiquitin ligase complex regulating NF-κB activity and apoptosis. *Nature* **471,** 637–641 (2011).
15. Tokunaga, F. *et al.* SHARPIN is a component of the NF-κB-activating linear ubiquitin chain assembly complex. *Nature* **471,** 633–636 (2011).
16. Niu, J., Shi, Y., Iwai, K. & Wu, Z. H. LUBAC regulates NF-κB activation upon genotoxic stress by promoting linear ubiquitination of NEMO. *EMBO J.* **30,** 3741–3753 (2011).
17. Nijman, S. M. *et al.* A genomic and functional inventory of deubiquitinating enzymes. *Cell* **123,** 773–786 (2005).
18. Mar, L., Rivkin, E., Kim, D. Y., Yu, J. Y. & Cordes, S. P. A genetic screen for mutations that affect cranial nerve development in the mouse. *J. Neurosci.* **25,** 11787–11795 (2005).
19. Hogan, K. A., Ambler, C. A., Chapman, D. L. & Bautch, V. L. The neural tube patterns vessels developmentally using the VEGF signaling pathway. *Development* **131,** 1503–1513 (2004).
20. Gondo, Y. Trends in large-scale mouse mutagenesis: from genetics to functional genomics. *Nature Rev. Genet.* **9,** 803–810 (2008).
21. Roca, C. & Adams, R. H. Regulation of vascular morphogenesis by Notch signaling. *Genes Dev.* **21,** 2511–2524 (2007).
22. Edelmann, M. J. *et al.* Structural basis and specificity of human otubain 1-mediated deubiquitination. *Biochem. J.* **418,** 379–390 (2009).
23. Juang, Y. C. *et al.* OTUB1 co-opts Lys48-linked ubiquitin recognition to suppress E2 enzyme function. *Mol. Cell* **45,** 384–397 (2012).
24. Wiener, R., Zhang, X., Wang, T. & Wolberger, C. The mechanism of OTUB1-mediated inhibition of ubiquitination. *Nature* **483,** 618–622 (2012).
25. Wang, T. *et al.* Evidence for bidentate substrate binding as the basis for the K48 linkage specificity of otubain 1. *J. Mol. Biol.* **386,** 1011–1023 (2009).
26. Matsumoto, M. L. *et al.* Engineering and structural characterization of a linear polyubiquitin-specific antibody. *J. Mol. Biol.* **418,** 134–144 (2012).
27. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437,** 1173–1178 (2005).
28. Maretto, S. *et al.* Mapping Wnt/β-catenin signaling during mouse development and in colorectal tumors. *Proc. Natl Acad. Sci. USA* **100,** 3299–3304 (2003).
29. HogenEsch, H. *et al.* A spontaneous mutation characterized by chronic proliferative dermatitis in C57BL mice. *Am. J. Pathol.* **143,** 972–982 (1993).
30. Seymour, R. E. *et al.* Spontaneous mutations in the mouse Sharpin gene result in multiorgan inflammation, immune system dysregulation and dermatitis. *Genes Immun.* **8,** 416–421 (2007).
31. Rahighi, S. *et al.* Specific recognition of linear ubiquitin chains by NEMO is important for NF-κB activation. *Cell* **136,** 1098–1109 (2009).
32. Mainardi, P. C. *et al.* The natural history of Cri du Chat Syndrome. A report from the Italian Register. *Eur. J. Med. Genet.* **49,** 363–383 (2006).
33. Mainardi, P. C. *et al.* Clinical and molecular characterisation of 80 patients with 5p deletion: genotype-phenotype correlation. *J. Med. Genet.* **38,** 151–158 (2001).
34. Zhang, X. *et al.* High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am. J. Hum. Genet.* **76,** 312–326 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Saunders and the Michigan University Transgenic Core facility for generating gumby BAC transgenic mice; K. Iwai for HOIL and HOIP constructs; K. Nakajima for NF-κB and AP2 reporter constructs; B. Alman and C. C. Hui for TOPGAL reporter mice; J. Woodgett for TOPFLASH, FOPFLASH and PRL vectors; T. Pawson for WNT3A expressing L-cells; M. Barrios-Ramos for help with fluorometry; J. Culotti and C. C. Hui for critical feedback; and I. Kourinov and staff at NE-CAT for collecting diffraction data set. B.R. holds the Canada research chair in proteomics and molecular medicine. A.-C.G. holds the Canada research chair in Functional Proteomics. F.S. holds the Canada research chair in Structural Principles of Signal Transduction. S.M.A. and T.S. were supported by CIHR predoctoral student fellowships. This work was supported by operating funds from the JSPS KAKENHI (grant numbers 15200032 to Y.G. and 21240043 to Y.G. and R.F.), and Canadian Institutes for Health Research to B.R. (MOP119289), A.-C.G. (MOP123433), F.S. (MOP57795) and S.P.C. (MOP 97966, IHO 94384 and MOP 111199).

Author Contributions E.R. designed genetic experiments with S.P.C., identified and confirmed the *gumby* causative mutation, performed expression analyses and characterized the *Gum*^{W96R} angiogenic phenotype; S.M.A. performed protein interaction assays and analysed the gumby–LUBAC connection; D.F.C. and Y.-C.J. performed X-ray crystallographic analyses and DUB assays; T.A.M. analysed angiogenesis and Wnt signalling; T.S. performed DUB chain profiling assay; H.H. and Y.-C.J. performed ITC analyses; Y.G. supervised R.F.; R.F. identified the *Gum*^{D336E} mutant; W.H.D. ran the mass spectrometry; A.-C.G. supervised W.H.D.; G.X. helped with SNP mapping; B.R. supervised T.S. and F.S. supervised D.C., H.H., Y-C.J. and together they designed experiments and wrote the manuscript; S.P.C. conceived and coordinated the project, designed and performed experiments, supervised E.R., S.M.A. and T.A.M., and wrote the manuscript.

Author Information The coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers 4KSJ (*apo*-gumby), 4KSK (gumby–ubiquitin complex) and 4KSL (gumby–linear diubiquitin complex). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.S. (sicheri@lunenfeld.ca) or S.P.C. (cordes@lunenfeld.ca).

METHODS

X-ray structure analyses, binding studies using ITC, deubiquitination assays, RNA and protein expression analyses, immunoprecipitations and Wnt reporter expression analyses *in vivo* and in cell culture were performed by standard methods.

Mouse husbandry. All mouse husbandry and handling was performed in conformity with the Canadian Council of Animal Care recommendations (AUP 0024a-00H). For all experiments, unless stated otherwise, a minimum of six *gumby* (Gum^{W96R}) and Gum^{D336E} homozygotes and six +/+ littermates were examined. Both mutations arose and were ultimately maintained on the C57BL6/J background.

For fine-mapping the *gumby* (Gum^{W96R}) mutation, heterozygous carriers and homozygous mutants were identified with D15Mit130 and D15Mit111. Recombinant embryos were scored with D15Mit252, D15Mit280, D15Mit18, D15Mit201 from http://www.informatics.jax.org/strains_SNPs.shtml and SNP markers rs3707949 and rs13482490 from http://www.ncbi.nlm.nih.gov/projects/SNP. Once identified, a tetra-primer ARMS-PCR assay was used to identify the Gum^{W96R} and Gum^{D336E} mutations essentially as described previously[35].

All primer sequences are listed in Supplementary Table 3.

Mice transgenic for bMQ-396D3 BAC (Chr.15: 27,544,413–27,629,541 bp) from AB2.2 embryonic stem cell DNA (129S7/SvEvBrd-$Hprt^{b-m2}$)[36] were generated by the Michigan University Transgenic Core into (C57BL/6 X SJL)F_2 females. The breeding strategy to produce Gum^{W96R}/Gum^{W96R};BAC transgenic mice is shown in Supplementary Fig. 3.

To assess canonical Wnt signalling, homozygous TOPGAL mice[37] were crossed with Gum^{W96R}/+ and Gum^{D336E}/+ mice. Gum^{W96R}/+;*TOPGAL*/+ mice and Gum^{D336E}/+;*TOPGAL*/+ mice were backcrossed to their respective heterozygous mutant backgrounds to obtain Gum^{W96R}/Gum^{W96R};*TOPGAL*/+ and Gum^{D336E}/Gum^{D336E};TOPGAL/+ embryos.

Sequencing. Coding sequences of candidate genes were amplified by PCR of exons and flanking introns from genomic DNA or cDNA from *gumby/gumby* E11.5 embryos. For expression constructs, wild-type gumby coding region was amplified from I.M.A.G.E. clone 4430188 and $gumby^{W96R}$ from Gum^{W96R}/Gum^{W96R} E10.5 embryonic cDNA using the Superscript first-strand synthesis system (Invitrogen). All clones were verified by sequencing.

Expression analyses. Northern blot analyses were performed using total RNA purified with Trizol (Invitrogen) at 20 μg per lane and resolved as described previously[38].

Whole-mount RNA *in situ* hybridization was performed, as previously described[39], using *gumby* cDNA region 459–1,059 bp, corresponding to AA153-352, cloned into SacI/EcoRI sites of pBluescript SKII (Stratagene). Templates for antisense probes were generated by XhoI digestion and transcription with T7 polymerase, and sense probes by ClaI digestion and transcription with T3 polymerase.

Anti-gumby polyclonal antibody. Glutathione *S*-transferase (GST)-tagged gumby protein (gumby–GST) was used to generate anti-gumby antibody in rabbits by the Laboratory Division of Comparative Medicine, University of Toronto. Antibody was purified with Affi-Gel15 coupled to maltose binding protein-tagged gumby fusion protein (MBP–gumby). Vectors used were pGEX-HTa (GE Healthcare) and pMAL-C2 vector (New England Biolabs).

Antibody specificity was tested by diluting in either 5% non-fat instant milk powder/TBST (immunoblot) or 10% normal goat serum/PBST (immunofluorescence) and incubating it with either GST–gumby or GST alone, immobilized to glutathione–Sepharose 4B beads (GE Healthcare). After 1.5 h of rocking at room temperature, beads were spun down, and the supernatant was used as a primary antibody (Supplementary Fig. 6).

Immunofluorescence. Cells, embryos and tissues were fixed in 4% paraformaldehyde (Sigma). Embryos and tissues were embedded in Tissue-Tek OCT (Miles laboratories) and sectioned at 16–20 μm. Samples were washed in PBS, blocked in PBST (PBS-0.3% Triton X-100) with 10% NGS for 2.5 h at room temperature, incubated with primary antibodies diluted in PBST-1% NGS overnight at 4 °C, washed in PBS, incubated with fluorescently labelled secondary antibodies for 1 h at room temperature, washed in PBS and mounted with Vectashield Mounting Medium with DAPI (Vector Laboratories). Whole-mount immunohistochemistry with anti-CD31 antibody (1:1,000) was performed as previously described[40].

Antibodies used were: rabbit anti-mouse gumby (1:500); mouse monoclonal anti-human smooth muscle actin (clone 1A4 DAKO, 1:1,000); monoclonal mouse anti-human desmin (clone D33 DAKO, 1:1,000); mouse anti-Flag M2 (Clone F1804 Sigma, 1:1,000); hamster anti-mouse CD31 (PECAM-1) (Millipore 1398Z, 1:500); chick anti-β-galactosidase (Abcam); anti-linear ubiquitin (Genentech, 1 μg ml^{-1}) antibodies. Secondary antibodies were goat anti-rat FITC, donkey anti-chick Cy3, donkey anti-hamster Cy2, donkey anti-rabbit Cy3 (all from Jackson Immunoresearch); donkey anti-mouse AlexaFluor 488, goat anti-mouse AlexaFluor 594, goat anti-rabbit AlexaFluor 488, goat anti-rabbit AlexaFluor 594 (all from Molecular Probes); goat anti-chick Alexa 488, goat anti-hamster Alexa 647 and goat anti-chick Alexa 488 (all from Invitrogen).

Immunofluorescence was visualized using a Nikon D-eclipse C1 confocal microscope system, Leica MZFIII microscope equipped with Qimaging 1300C digital camera or Leica spinning disc confocal microscope and analysed with Adobe Photoshop and Volocity or ImageJ.

Anti-β-gal immunofluorescence. Immunofluorescence was performed on matched sections blocked for 1 h with 5% NGS, 5% normal donkey serum (NDS), 1% blocking solution (prepared from a 10% solution in maleate buffer), 20 mM $MgCl_2$, 0.3% Tween-20 in PBS. Chick anti-β-galactosidase, Cy3-anti-chicken and Cy2-anti-hamster were each pre-incubated for 1 h at room temperature with 3 mg mouse embryo powder in 2% NGS, 2% NDS, 1% blocking solution, 20 mM $MgCl_2$, 0.3% Tween-20 in PBS, centrifuged at 10,000*g* for 10 min at 4 °C, then applied to sections at appropriate dilution. Three lots of chick anti-β-gal antibody were tested to identify one optimal preparation. Sections were incubated with anti-β-gal and hamster anti-PECAM-1 antibodies for 14–18 h at 4 °C, washed five times with PBS/0.2% Tween. Secondary antibodies were applied at 1:400 dilutions for 45 min. Sections were washed in PBS three times and mounted as usual. Using serial dilutions ranging from 1:100 to 1:4,000, we determined that immunofluorescent anti-β-gal signal was saturated at 1:200 to 1:500, not detectable at 1:2,000, and present at intermediate levels at 1:1,000 dilution on +/+;*TOPGAL*/+ sections.

To establish gumby and β-gal co-localization in endothelial cells within ISVs, we used the 1:500 anti-β-gal dilution, analysed more than three animals per genotype and found complete co-localization in ISVs. For quantification purposes, we counted >100 cells from >10 ISVs per genotype. Gumby and β-gal expression in other embryonic endothelial cells is less uniform. Given the phenotypes we observed in blood vessels within the neural tissue, we focused analyses on these by sampling 394 endothelial cells in 20 transverse sections through an E10.5 +/+;*TOPGAL*/+ embryo.

We used the 1:1,000 dilution in subsequent quantification experiments. Experiments were performed in parallel on a minimum of six sections from three animals per genotype. Images were collected as 3.5-μm stacks under identically set conditions for all samples. For quantification, at least 15 ISVs located between the fore- and hindlimbs were sampled per genotype. Immunofluorescence was analysed using ImageJ software[41]. PECAM-positive endothelial cells in ISVs were traced as a group and the mean fluorescent intensity of ISVs (that is, of endothelial cells within ISVs) was measured. A Kruskal–Wallis test ($k = 3$) based on ranks was performed on the data, as they did not meet the parametric assumptions regarding normality and equivalence of variance. Effect of gumby genotype on anti-β-gal mean fluorescence intensity was highly significant ($H = 43.69$; $P < 0.0001$).

Immunoblot analysis. Transfected cells and mouse tissues were lysed in RIPA buffer (1% NP40, 0.5% sodium deoxycholate, 0.1% SDS, in PBS pH 7.4 with protease inhibitor cocktail tablet (Roche)). SDS–polyacrylamide gel electrophoresis was performed with 10–15 μg of total lysate and immunoblot analysis was performed as described. To control for loading, membranes were stripped and probed for tubulin.

Antibodies used were rabbit gumby (1:5,000); mouse tubulin (Clone DM1A Sigma, 1:2000); goat HRP-IgG (anti-rabbit, rat, human or mouse) (Jackson Immunoresearch, 1:10,000); mouse Flag (Sigma M2, 1:1,000); rat HA (Roche, 1:1,000); mouse Myc (Santa Cruz Biotechnology, 1:500); human linear ubiquitin (Genentech, 1:1,000) and rabbit pan-ubiquitin (Dako, 1:1,000)

Immunoprecipitation. Human HEK293T cells were transfected using Effectene transfection reagent (Qiagen). Cells were lysed 18–24 h after transfection in 50 mM Tris pH 7.4, 100 mM EDTA, 150 mM NaCl and 0.5% Triton X-100 and protease inhibitor cocktail (Roche). Aliquots containing 400 μg total protein were incubated for 1 h with 1 μg Flag antibody (Mouse M2, Sigma), 2 μg Myc antibody (mouse, Santa Cruz) or 2 μg HA antibody (rat, high affinity, Roche). 20 μl of Protein A/G plus agarose beads (Santa Cruz) were added and co-immunoprecipitation was performed as in ref. 42. For DVL2 immunoprecipitation, 4 μM MG132 was also added to media 18 h after transfection and during lysis and immunoprecipitation. Samples were analysed by immunoblot.

***In vitro* luciferase assays.** Wnt signalling was assayed in HEK293T cells co-transfected with various expression constructs and with TOPFLASH or FOPFLASH reporters[43] and the normalizing transfection efficiency control PRL vector using Effectene (Qiagen), incubated for 18–20 h and then stimulated for 4 h with WNT3A conditioned or control media. Similarly, NF-κB-dependent transcription was assayed with a reporter containing six NF-κB binding sites and a control plasmid with six AP1 binding sites that drive firefly luciferase expression in Dual luciferase assays (Promega)[15]. Values were normalized relative to Flag–pcDNA3.1. Constructs did not activate the FOPFLASH or AP1 reporters. At least three independent experiments were performed with samples in triplicate for each construct.

Analysis of linear ubiquitinated proteins. Analyses of linear ubiquitinated proteins by immunoblot and immunofluorescence were performed as described previously[26]. For embryonic lysates, embryos were collected at E10.5 and washed twice with PBS. Individual embryos and 293T cells were lysed in 8 M urea, 50 mM Tris pH 7.5, 25 mM NaCl, 2 mM EDTA, 2 mM *N*-ethylmaleimide and complete mini protease inhibitors (Roche) and were disrupted by passing through an 18 gauge needle.

Protein expression and purification. Human FAM105B residues 79–352 (gumby^{R79}) corresponding to the catalytic domain and residues 55–352 (gumby^{M55}) corresponding to the catalytic domain with an additional 24-residue N-terminal extension, were amplified by PCR from a full-length cDNA clone (IMAGE: 4430188) and cloned into pGEX-TEV downstream of GST and pProEx-TEV downstream of His$_6$. W96R, C129A, C129S and D336E mutations were introduced using standard techniques and all clones were sequence verified. Linear diubiquitin was generated by PCR amplification of residues 1–152 from Ubc (IMAGE:4076286) and cloned into pProEx.

Native and selenomethionine-labelled His$_6$-tagged gumby proteins, ubiquitin and linear di-ubiquitin were expressed in *Escherichia coli* BL21(DE3) Codon+ (Agilent Technologies) and purified using a HiTrap nickel chelating HP column (GE Healthcare) using standard laboratory protocols. Tag-free proteins in 20 mM HEPES pH 7.5, 100 mM NaCl, 5 mM β-mercaptoethanol were concentrated to 10–25 mg ml^{-1} and stored frozen at −80 °C.

5-Iodoacetamidofluorescein (5-IAF, Molecular Probes)-labelled linear diubiquitin was generated by incubating 100 μM purified diUb^{Cys0} (modified by inclusion of a non-native cysteine residue at the ubiquitin amino terminus) with 500 μM of 5-IAF in a buffer containing 20 mM HEPES pH 7.5, 100 mM NaCl, and 1 mM dithiothreitol (DTT) for 3 h at 20 °C followed by size exclusion chromatography.

Purified USP2, USP5 and USP21 were obtained as described previously[44].

Structure determination by X-ray crystallography. Crystals of gumby^{R79} were grown at 20 °C by mixing equal volumes of 600 μM protein with solution containing 100 mM MES pH 6.0, 200 mM MgCl$_2$, 19% PEG3350 and cryo-protected with 20% glycerol. Diffraction to 1.60 Å resolution ($\lambda = 0.97917$) was collected at NE-CAT 24-ID-C beamline at 100K and processed using HKL2000. Phased electron density maps from nine seleno-methionine positions were generated by SHELX. Protein structure was built using Arp/Warp and refined using Phenix and Coot. 96.4%, 3.6% and 0% of the residues resided in the most favoured, allowed and disallowed regions of the Ramachandran plot, respectively (Procheck). Crystals of an equimolar mixture of gumby$^{M55\text{-}C129S}$ and ubiquitin (900 μM final concentration) were grown at 20 °C by mixing equal volumes of protein with solution containing 100 mM Bis Tris pH 5.5, 200 mM ammonium sulphate, 20% PEG3350 and cryo-protected with 20% glycerol. Diffraction data was collected at 100K at a home source ($\lambda = 1.5418$) with Rigaku Saturn944+ detector and processed using HKL3000. Crystals of an equimolar complex of gumby$^{R79\text{-}C129A}$ and linear di-ubiquitin were grown at 20 °C by mixing equal volumes of protein with a solution of 100 mM acetate pH 5.5, 100 mM CaCl$_2$, 100 mM glycine, 2.5 M sodium formate and 24% PEG 3350 and cryoprotected with Paratone-N. Diffraction data was collected at NE-CAT 24-ID-C beam line ($\lambda = 0.97917$) at 100K and processed using HKL2000. The gumby–ubiquitin and gumby–linear diubiquitin co-structures were determined by molecular replacement using Phaser with gumby^{R79} and ubiquitin as search models. 94.8%, 4.7% and 0.5% of residues in the gumby–ubiquitin complex and 96.8%, 2.4% and 0.8% of residues in the gumby–diubiquitin complex reside in most favoured, allowed and disallowed regions, respectively. Statistics for data collection, phasing and structure refinement are presented in Supplementary Table 1.

Deubiquitination assay. Substrate specificity analysis of gumby and OTUB1 shown in Fig. 3b was determined by incubating 1 μg of the indicated diubiquitin substrates (Boston Biochem) with 1 μg of the indicated DUB in a 20 μl reaction with 20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM DTT at 37 °C for 18 h. Reactions were stopped by the addition of 6× Laemmli buffer. Reaction products were subjected to 4–12% PAGE and stained with Coomassie brilliant blue.

For enzyme titration experiments in Supplementary Fig. 8a and e, 1 μg of linear diubiquitin or K48-linked diubiquitin was incubated with the indicated concentration of gumby or OTUB1 in 20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM DTT at 37 °C for 30 min in a 10-μl volume. Reactions were stopped by the addition of 6× Laemmli buffer, subjected to 15% PAGE and gel stained with Coomassie brilliant blue.

Kinetic analyses of gumby^{M55}, gumby^{R79}, gumby^{D336E} and gumby^{W96R} DUB activity were performed by determining the initial reaction rates from a series of diubiquitin cleavage reactions containing 0.08 to 150 μM of fluorescein-labelled linear diubiquitin substrate. Reactions were performed at 37 °C in 20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM DTT by mixing gumby (0.5 nM gumby^{M55}, 0.4 nM gumby^{R79}, 30 nM gumby^{D336E} or 25 μM gumby^{W96R}) and fluorescein-labelled linear diubiquitin at 0.08, 0.4, 2, 10, 25, 50, 100 or 150 μM concentration. 35 μl aliquots were obtained at various time points from a 600 μl reaction volume, stopped by the addition of 6× Laemmli buffer, and followed by 15% PAGE analysis. The gel was imaged with a Typhoon FLA 9500 (GE Healthcare) using the Alexa-Fluor 488nM protocol. Bands on the gel corresponding to fluorescein-labelled monoubiquitin and standards of fluorescein-labelled diubiquitin were used to quantify the amount of reaction product at each time point using ImageQuant (GE Healthcare). Initial rates for each reaction performed in duplicate were calculated from the linear portions of each reaction profile and fitted to a nonlinear Michaelis–Menten equation using Prism v5 (GraphPad Software Inc.) to calculate K_m, V_{max} and k_{cat} values. (See Supplementary Microsoft Excel spreadsheet for replicate data.)

For ubiquitin-vinyl sulphone inhibition assays shown in Supplementary Fig. 8b, 0.6 μM gumby, USP2, USP5 and USP21 were incubated with indicated amounts of ubiquitin-vinyl sulphone (Boston Biochem) for 30 min at 37 °C and subsequently used in a DUB assay as described above. Ub–AMC assays shown in Supplementary Fig. 8c were carried out with 5, 50 and 500 nM of the indicated DUBs and 600 nM of Ub–AMC (Boston Biochem) in a buffer containing 50 mM HEPES pH 7.5, 100 mM NaCl, 0.03% Brij-35, and 0.1 mg ml^{-1} BSA for 2 h at 30 °C. Fluorescence was measured using an EnSpire 2300 plate reader (PerkinElmer) at 380 nm excitation and 480 nm emission wavelengths.

Isothermal titration calorimetry. Calorimetric titrations were carried out on a Microcal VP-ITC titration calorimeter. Protein samples were prepared in 20 mM HEPES, pH 7.5, 300 mM NaCl, 1 mM DTT. 0.2–0.5 mM linear diubiquitin within the syringe was titrated into 20–50 μM concentrations of gumby^{C129S} or gumby$^{C129S/W96R}$ proteins residing in the sample cell. Experiments were carried out minimally in triplicate at 30 °C. Data analysis was performed using Origin software (Microcal) (See Supplementary Microsoft Excel spreadsheet for replicate data).

Mass spectrometric analysis. For mass spectrometry, generation of stable inducible cell lines, induction of protein expression and affinity purification on Flag M2 magnetic beads was performed as described[45]. Proteins digested with trypsin on beads were loaded onto reversed-phase capillary columns and analysed by LC-MS/MS on a LTQ mass spectrometer, as described previously[46]. RAW files were saved in our local interaction proteomics LIMS, ProHits[47], searched with Mascot version 2.3 (Matrix Science) against the human and adenovirus complements of the RefSeq protein database (version 45; 34604 entries), enabling one missed cleavage site, and Met oxidation and Asn/Gln deamidation as variable modifications. Search results were further analysed within ProHits[45] using the SAINT statistical tool[48,49]. SAINT analysis was done using two biological replicates per bait. Bait protein samples were analysed alongside 8 negative control runs (Flag alone), using the following parameters: 8 negative controls compressed into 6 (nControl:6); nburn: 2000, niter: 5000, lowMode: 0; minFold: 1, normalize: 1 (ref. 48). After SAINT analysis, results were further filtered to show only those proteins identified in with an average SAINT probability of 0.9, at least 2 unique peptides, and with a frequency of identification less than in 10% of all the samples annotated in a database consisting of >2,000 Flag AP–MS experiments.

35. Ye, S., Dhillon, S., Ke, X., Collins, A. R. & Day, I. N. An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.* **29,** e88 (2001).
36. Adams, D. J. *et al.* A genome-wide, end-sequenced 129Sv BAC library resource for targeting vector construction. *Genomics* **86,** 753–758 (2005).
37. DasGupta, R. & Fuchs, E. Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation. *Development* **126,** 4557–4568 (1999).
38. Ambulos, N. P. Jr, Duvall, E. J. & Lovett, P. S. Method for blot-hybridization analysis of mRNA molecules from *Bacillus subtilis*. *Gene* **51,** 281–286 (1987).
39. Kim, F. A. *et al.* The vHNF1 homeodomain protein establishes early rhombomere identity by direct regulation of Kreisler expression. *Mech. Dev.* **122,** 1300–1309 (2005).
40. Vecchi, A. *et al.* Monoclonal antibodies specific for endothelial cells of mouse blood vessels. Their application in the identification of adult and embryonic endothelium. *Eur. J. Cell Biol.* **63,** 247–254 (1994).
41. Abramoff, M. D., Magalhaes, P. J. & Ram, S. J. Image Processing with ImageJ. *Biophotonics Intl* **11,** 36–42 (2004).
42. Torban, E., Wang, H. J., Groulx, N. & Gros, P. Independent mutations in mouse Vangl2 that cause neural tube defects in looptail mice impair interaction with members of the Dishevelled family. *J. Biol. Chem.* **279,** 52703–52713 (2004).
43. Veeman, M. T., Axelrod, J. D. & Moon, R. T. A second canon. Functions and mechanisms of β-catenin-independent Wnt signaling. *Dev. Cell* **5,** 367–377 (2003).
44. Ernst, A. *et al.* A strategy for modulation of enzymes in the ubiquitin system. *Science* **339,** 590–595 (2013).
45. Kean, M. J., Couzens, A. L. & Gingras, A. C. Mass spectrometry approaches to study mammalian kinase and phosphatase associated proteins. *Methods* **57,** 400–408 (2012).

46. Dunham, W. H. *et al.* A cost–benefit analysis of multidimensional fractionation of affinity purification-mass spectrometry samples. *Proteomics* **11,** 2603–2612 (2011).
47. Liu, G. *et al.* ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nature Biotechnol.* **28,** 1015–1017 (2010).
48. Choi, H. *et al.* Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr. Protoc. Bioinform.* Ch. 8, Unit 8.15 (2012).
49. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods* **8,** 70–73 (2011).

doi:10.1038/nature12204

RAS–MAPK–MSK1 pathway modulates ataxin 1 protein levels and toxicity in SCA1

Jeehye Park[1,2,3]*, Ismael Al-Ramahi[1,2]*, Qiumin Tan[1,2,3], Nissa Mollema[4,5], Javier R. Diaz-Garcia[1,2], Tatiana Gallego-Flores[1,2], Hsiang-Chih Lu[2,6], Sarita Lagalwar[4,5], Lisa Duvick[4,5], Hyojin Kang[1,2]†, Yoontae Lee[1,2,3]†, Paymaan Jafar-Nejad[1,2], Layal S. Sayegh[1,2], Ronald Richman[1,2,3], Xiuyun Liu[1,2,3], Yan Gao[1,2], Chad A. Shaw[1], J. Simon C. Arthur[7], Harry T. Orr[4,5], Thomas F. Westbrook[1,6,8], Juan Botas[1,2] & Huda Y. Zoghbi[1,2,3,6]

Many neurodegenerative disorders, such as Alzheimer's, Parkinson's and polyglutamine diseases, share a common pathogenic mechanism: the abnormal accumulation of disease-causing proteins, due to either the mutant protein's resistance to degradation or overexpression of the wild-type protein. We have developed a strategy to identify therapeutic entry points for such neurodegenerative disorders by screening for genetic networks that influence the levels of disease-driving proteins. We applied this approach, which integrates parallel cell-based and *Drosophila* genetic screens, to spinocerebellar ataxia type 1 (SCA1), a disease caused by expansion of a polyglutamine tract in ataxin 1 (ATXN1). Our approach revealed that downregulation of several components of the RAS–MAPK–MSK1 pathway decreases ATXN1 levels and suppresses neurodegeneration in *Drosophila* and mice. Importantly, pharmacological inhibitors of components of this pathway also decrease ATXN1 levels, suggesting that these components represent new therapeutic targets in mitigating SCA1. Collectively, these data reveal new therapeutic entry points for SCA1 and provide a proof-of-principle for tackling other classes of intractable neurodegenerative diseases.

The increasing prevalence of neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease in the ageing population is one of our most pressing public health issues[1,2]. There is no disease-modifying treatment for most of these disorders, and therapies that do exist target the late stages of disease, when pathogenesis is already advanced and the symptomatic relief is partial and short-lived[3,4]. It would clearly be preferable to delay the onset of symptoms, slow disease progression or reverse the course of the disease, but to do this probably requires us to address earlier stages of pathogenesis.

Fortunately, two decades of research on inherited neurodegenerative diseases (Huntington's disease, the spinocerebellar ataxias and inherited forms of Alzheimer's disease and Parkinson's disease) have revealed certain pathogenic commonalities at the molecular level[5]. For instance, although the respective disease-causing proteins (HTT, ataxins, APP and α-synuclein) and neuronal populations vulnerable to their toxicity are distinct in each disorder[6–8], each of these diseases involves the toxic accumulation of the mutant protein; in some cases too much of the wild-type protein is also toxic[9–11]. Importantly, animal models have shown that decreasing the accumulation of the normal or mutant proteins, usually by genetic manipulation, can reverse disease phenotypes[12–18]. We therefore set out to identify druggable targets that modulate the level of a neurodegenerative disease-causing protein: glutamine-expanded ATXN1, which is mutated in SCA1.

To exploit the unbiased, functional nature of genetic screens and compensate for the weaknesses inherent in any given model system, we developed a cross-species strategy: complementary forward genetic screens that targeted the levels of glutamine-expanded human ATXN1 in a SCA1 *Drosophila* model and in a human cell model for SCA1. We chose to use modulation of ATXN1 levels as a proof-of-principle for this strategy for three reasons: (1) the severity of neurodegeneration in SCA1 correlates with the levels of the mutant ATXN1 protein; (2) overexpression of wild-type ATXN1 leads to neurodegeneration; and (3) the pathogenic mechanisms underlying SCA1 are well characterized[19]. Our genetic screening approach revealed that multiple components of the RAS–MAPK–MSK1 pathway influence ATXN1 levels in *Drosophila* and human cells. We then validated these results using SCA1 mouse models and further found that pharmacological targeting of this pathway also reduces ATXN1 levels.

Strategy to identify regulators of ATXN1 levels

To identify regulators of ATXN1 levels in human cells of neural lineage, we engineered a human medulloblastoma-derived cell line with a transgene encoding a glutamine-expanded ATXN1 fused with monomeric red fluorescent protein (mRFP–ATXN1(82Q)). To distinguish modifiers regulating ATXN1 protein levels from those affecting transgene transcription, we placed an internal ribosomal entry site (IRES) followed by yellow fluorescent protein (YFP) downstream of the ATXN1 fusion protein (mRFP–ATXN1(82Q)–IRES–YFP). The ratio of mRFP to YFP fluorescence by flow cytometric analysis thus serves as a read-out for genes that regulate mRFP–ATXN1(82Q) protein levels while controlling for fluctuations in transcription of the transgene (Fig. 1a).

We focused on interrogating human kinases, because (1) phosphorylation is crucial to ATXN1 toxicity[20]; and (2) many kinases are targeted (or amenable to targeting by) pharmacological agents[21]. We used this cell system to test individually the effects of short interfering RNAs

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [2]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, Texas 77030, USA. [3]Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas 77030, USA. [4]Institute for Translational Neuroscience, University of Minnesota, Minneapolis, Minnesota 55455, USA. [5]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, Minnesota 55455, USA. [6]Program in Developmental Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [7]MRC Protein Phosphorylation Unit, School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK. [8]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. †Present addresses: Supercomputing Center, Korea Institute of Science and Technology Information, Daejeon 305-806, South Korea (H.K.); Department of Life Science, Pohang University of Science and Technology, Pohang, Kyungbuk 790-784, South Korea (Y.L.).

*These authors contributed equally to this work.

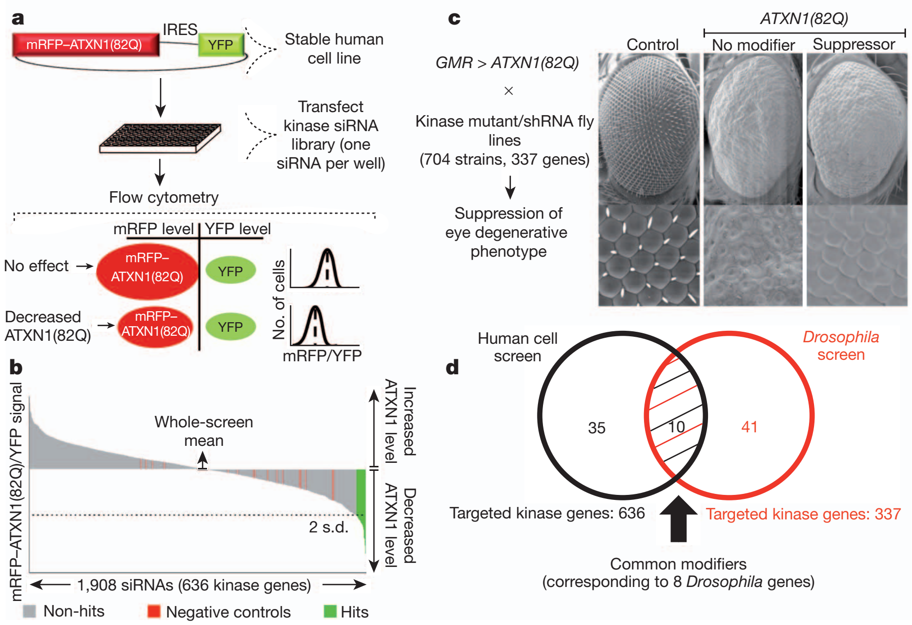


Figure 1 | Integrative genetic screen identifies regulators of ATXN1(82Q) stability. a, Forward genetic screen strategy for modifiers of ATXN1(82Q) levels. mRFP–ATXN1(82Q)–IRES–YFP-expressing cells are used to assess the abundance of ATXN1(82Q) by monitoring the mRFP–ATXN1(82Q) to YFP fluorescence ratio. siRNAs that reduce ATXN1(82Q) protein lead to a decrease in mRFP–ATXN1(82Q) but not YFP. **b**, Primary screen result shows the effect on the average mRFP–ATXN1(82Q)/YFP ratio per kinase siRNAs tested. **c**, Schematic of the *Drosophila* screen for suppressors of ATXN1(82Q). Suppressor example is *lic* (homologue of human MEK3). **d**, Venn diagram of the identified gene candidates.

(siRNAs) targeting every known human kinase and kinase-like gene (1,908 siRNAs, 636 genes; tests performed in triplicate) on ATXN1 levels. We selected for further study 181 siRNAs that reduced the ratio of mRFP to YFP fluorescence by two standard deviations from the screen-wide mean (Fig. 1b). siRNAs from these primary hits were transfected independently in ATXN1(82Q) cells for confirmation and were tested in parallel with a control reporter line (cells encoding mRFP–IRES–YFP lacking the ATXN1 fusion) to eliminate false-positive hits that targeted mRFP. These analyses validated 50 siRNA candidates (corresponding to 45 genes) that reduce ATXN1(82Q) levels (Supplementary Table 1).

In parallel, we performed a genetic screen in a *Drosophila* SCA1 model expressing human ATXN1(82Q), in which fruitflies develop an external eye phenotype in response to ATXN1 toxicity[22] (Fig. 1c). We screened a total of 704 alleles (including inducible short hairpin RNAs (shRNAs) and loss-of-function mutations) corresponding to 337 kinase-encoding *Drosophila* genes for those that would modulate mutant ATXN1 levels. We then performed retinal sections to determine whether these modulators of the external eye phenotype also improved photoreceptor integrity. The morphological and histological phenotype assessment yielded 51 alleles (corresponding to 49 genes) that suppressed ATXN1(82Q) toxicity *in vivo* (Supplementary Table 2).

These *Drosophila* and human-cell-based screens revealed ten human modifier genes that reduce both ATXN1 levels and ATXN1-induced toxicity (*IGF1R*, *ULK3*, *WNK4*, *BUB1*, *MSK1* (also known as *RPS6KA5*), *MEK2* (also known as *MAP2K2*), *MEK3* (also known as *MAP2K3*), *MEK6* (also known as *MAP2K6*), *ERK1* (also known as *MAPK3*) and *ERK2* (also known as *MAPK1*)) and that correspond to eight modifiers in *Drosophila* (*CG3837* (also known as *Sdr*), *CG8866*, *CG7177* (also known as *Wnk*), *BubR1*, *JIL-1*, *Dsor1*, *lic* and *rl*) (Figs 1c, d, 2a–c and Supplementary Fig. 1). We selected these genes for further characterization.

Reducing MAPK signalling decreases ATXN1

Network analysis of the hits from human cells and *Drosophila* revealed that the MAPK cascade is the most enriched pathway in each screen (Supplementary Table 3). In addition, analysis of the ten common genes shows that the MAPK cascade is the most highly represented signalling pathway. In fact, six of the ten shared hits (ERK1, ERK2, MEK2, MEK3, MEK6 and MSK1) are canonical components of this pathway, and WNK4 and IGF1R are known to regulate it[23–25] (Fig. 2e, Supplementary Table 4 and Supplementary Fig. 1d). This pathway is a component of a cell-signalling cascade that also includes RAS, RAF and MEK and that can be activated by a variety of mitogens, growth factors and receptor tyrosine kinases[26,27] (Supplementary Fig. 1d). To validate the effects of the MAPK pathway hits in the central nervous system of an intact animal, we used a motor performance assay that measures climbing ability in fruitflies. Expression of ATXN1(82Q) in the *Drosophila* central nervous system leads to progressive impairment in motor performance[28]. We found that decreased levels of the *Drosophila* homologues of MEK (DSOR1), ERK1/2 (RL) and MSK1 (JIL-1) suppressed ATXN1(82Q)-induced motor deficits and improved lifespan in ATXN1(82Q) flies (Fig. 2d).

We next sought to determine whether other upstream components of the MAPK pathway modulate ATXN1 toxicity. We found that decreased levels of the *Drosophila* homologues of farnesyltransferase (FNTB), GRB2, SOS, HRAS, RRAS and RAF suppress the ATXN1(82Q)-induced external eye phenotype as well as photoreceptor degeneration (Fig. 3a and Supplementary Fig. 2a, b). Decreasing the levels of the *Drosophila* homologues of FNTA, SOS, HRAS and RAF improved the ATXN1(82Q)-induced motor and lifespan phenotypes (Fig. 3b and Supplementary Fig. 3). In addition, activation of MAPKs via a constitutively active form of *Drosophila* RAS (dRAS-V12) exacerbated the ATXN1(82Q)-induced eye degeneration (Supplementary Fig. 4a).

Prompted by this genetic evidence that the RAS–MAPK–MSK pathway governs ATXN1(82Q)-induced phenotypes, we investigated the ability of this pathway to modulate ATXN1(82Q) protein levels. We found that reducing the levels of *HRAS* and *FNTA* led to a decrease in ATXN1 levels in human cells (Fig. 3c, d, Supplementary Fig. 2c and Supplementary Fig. 5). Furthermore, we confirmed that decreasing the levels of any of the two functionally conserved RAS homologues in *Drosophila* (dRAS (also known as RAS85D) and RAS64B) also decreased the levels of ATXN1(82Q) *in vivo* (Fig. 3e). Reducing the levels of the RAS/MAPK genes also decreased ATXN1(30Q) levels in cells and suppressed ATXN1(30Q)-induced toxicity in *Drosophila* (Supplementary Fig. 6).

In summary, the genetic screen in cells and fruitflies unveiled the RAS/MAPK pathway as critical for modulating ATXN1 levels (Fig. 3f).

MSK1 stabilizes ATXN1 by phosphorylating S776

Building on our previous discovery that ATXN1 levels are highly sensitive to phosphorylation at residue S776 (mutation of serine to alanine decreases both ATXN1 stability and toxicity)[20,29], we investigated whether the MAPK pathway kinases (MEK, ERK and MSK1)

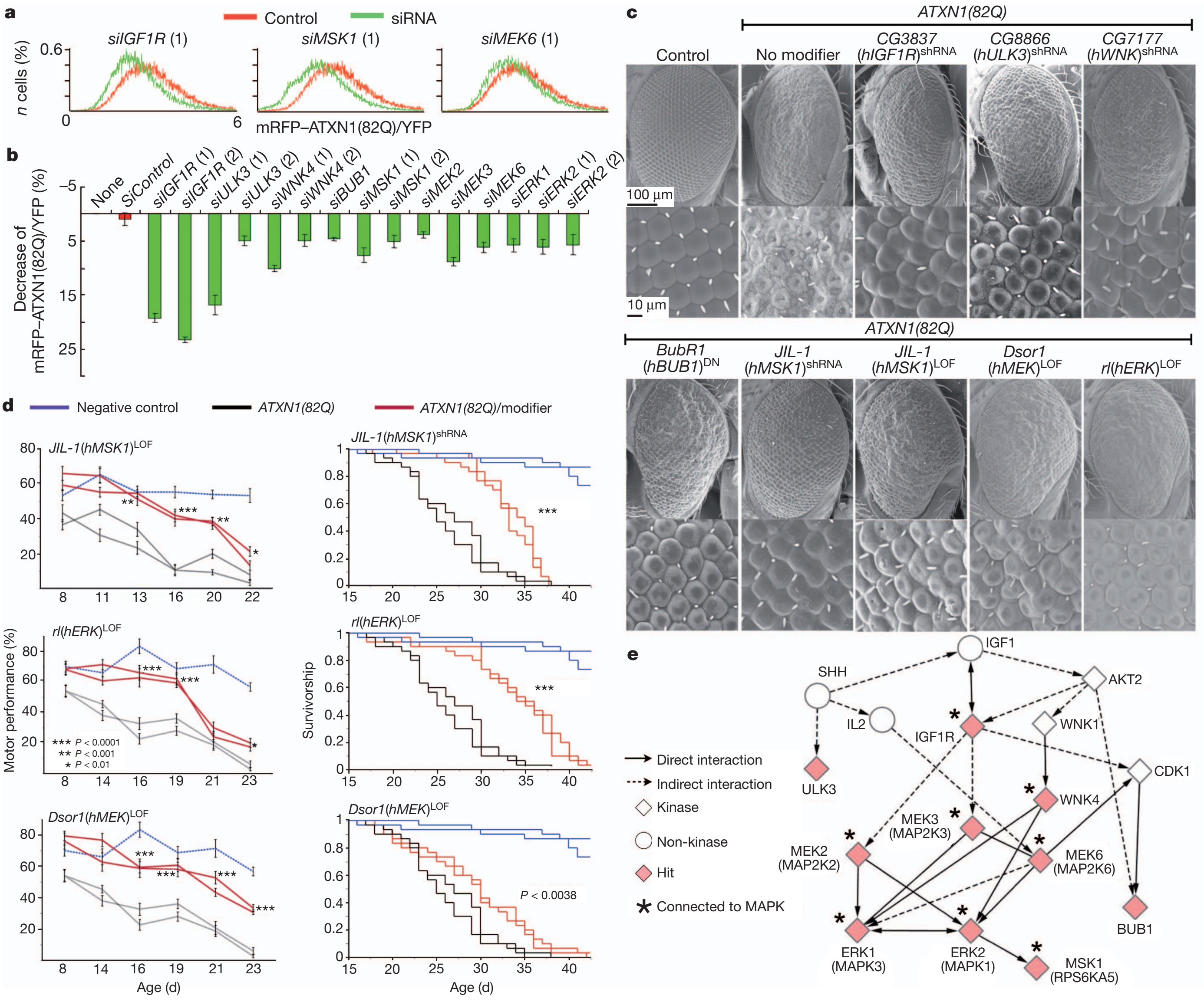


Figure 2 | Modifiers shared between cell-based and *in vivo* screens. **a**, Histograms show the distribution of mRFP–ATXN1(82Q)/YFP ratio abundance in cells treated with indicated siRNA compared to control. *n*, number. **b**, Average mRFP–ATXN1(82Q)/YFP ratio change upon treatment with siRNAs hits (error bars: s.e.m. from triplicates, $P < 0.05$). **c**, Scanning electron microscopy images of ATXN1(82Q) suppressors in *Drosophila* caused by reduced level of candidate genes in **b**. DN, dominant negative; LOF, loss of function. Names of human (h) homologues are noted in parentheses. **d**, Decreased levels of *Drosophila* MSK1, ERK and MEK homologues suppress motor impairment and improve survival. Error bars, s.e.m. $*P < 0.1$, $**P < 0.001$, $***P < 0.0001$. **e**, Pathway analysis showing eight of the ten modifiers connected to MAPK pathway. Diamonds represent modifiers common to both screens.

recovered in our screens phosphorylate ATXN1 at S776. This residue falls within the consensus phosphorylation sequence RXXS (in which X denotes any residue) that is conserved in ATXN1 homologues (Fig. 4a). Most MAPK proteins are known to prefer PX(S/T)P (in which brackets denote alternative residues), so we excluded MEK and ERK. Instead, we focused on MSK1, which is known to prefer the consensus sequence RXXS[30].

To determine whether MSK1 directly phosphorylates ATXN1 at S776, we conducted *in vitro* kinase assays with purified MSK1 and glutathione-*S*-transferase–ATXN1(82Q) (GST–ATXN1(82Q)). We immunoblotted with anti-phospho-S776 ATXN1 antibody[20] and found that the antibody detected robust phospho-S776 ATXN1 signal upon addition of MSK1 (Fig. 4b). Furthermore, MSK1 can phosphorylate both the mutant GST–ATXN1(82Q) and the wild-type GST–ATXN1(30Q) (Fig. 4b). It is noteworthy that RSK1 (also known as RPS6KA1) and RSK2 (also known as RPS6KA3), two other kinases that share the RXXS consensus sequence and are also downstream of the MAPK pathway[30], failed to phosphorylate the S776 site (Supplementary Fig. 7a–c).

We next investigated whether the observed effect of MSK1 on ATXN1 stability is mediated by S776 phosphorylation *in vivo*. We found increased levels of ATXN1(82Q) upon coexpression with Msk1 in mouse Neuro2A cells (Fig. 4c). This increase was dependent on S776 phosphorylation, because Msk1 did not increase the levels of ATXN1(82Q, S776A) (Fig. 4c). We also performed cerebellar fractionation assays from wild-type mouse brains and examined whether Msk1 was enriched in the fractions containing S776 phosphorylation activity. As shown in Supplementary Fig. 8a, Msk1 was enriched in the fractions containing peak S776-phosphorylating activity. To confirm whether endogenous Msk1 from neural lineages contributes to ATXN1 phosphorylation, we immunodepleted Msk1 from mouse cerebellar extracts and found that reduction of Msk1 significantly

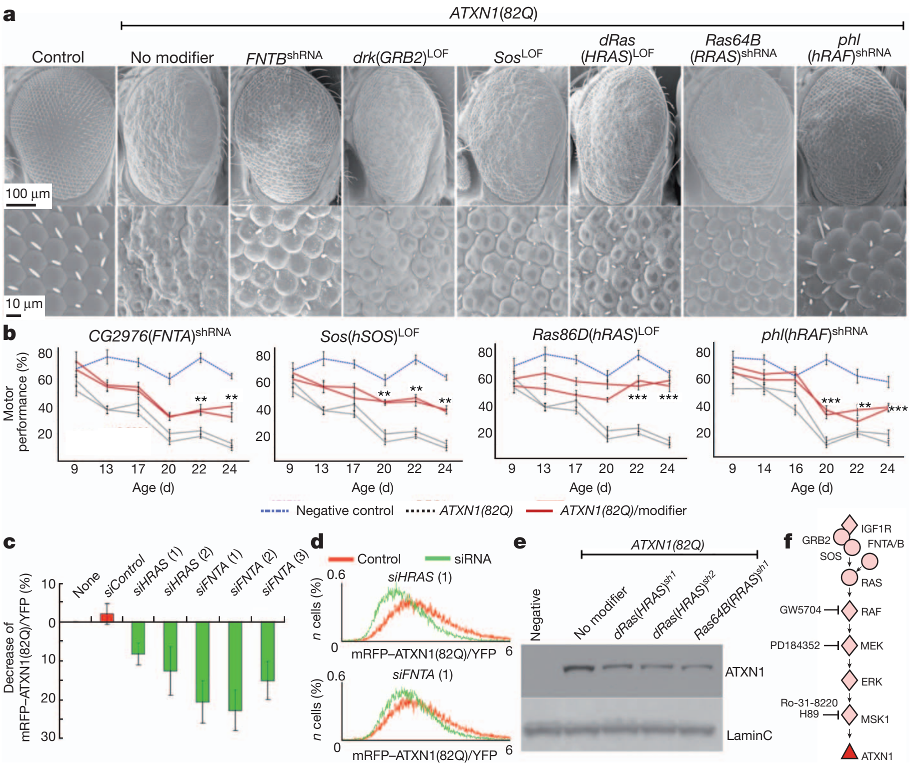


Figure 3 | Upstream MAPK pathway components regulate ATXN1 toxicity and levels. **a**, Decreased levels of *Drosophila* homologues of the MAPK upstream components suppress ATXN1(82Q)-induced eye degeneration. **b**, Decreased levels in the *Drosophila* homologues of FNTA, SOS, HRAS and RAF suppress ATXN1(82Q)-induced motor impairments. $**P < 0.001$, $***P < 0.0001$. **c**, Change in mRFP–ATXN1(82Q)/YFP average fluorescence ratio upon treatment with indicated siRNAs. Error bars, s.e.m. from triplicates, $P < 0.05$. **d**, mRFP–ATXN1(82Q)/YFP ratio distribution in siRNA-treated whole-cell populations compared to control. **e**, Decreased levels of ATXN1(82Q) induced by RAS shRNA in *Drosophila*. **f**, MAPK pathway showing the ATXN1(82Q) modifiers and Fig. 5 inhibitors. Diamonds represent modifiers from the screen.

decreased S776 phosphorylation of bacterially purified ATXN1 compared to non-depleted extracts (Fig. 4d and Supplementary Fig. 7d).

Having established the role of MSK1 in phosphorylating a serine critical for modulating ATXN1 stability, we investigated the effects of MSK1 on ATXN1 level in *Drosophila*. We found that reducing the level of the *Drosophila* homologue of MSK1 decreased the steady-state level of ATXN1(82Q) *in vivo* (Fig. 4e).

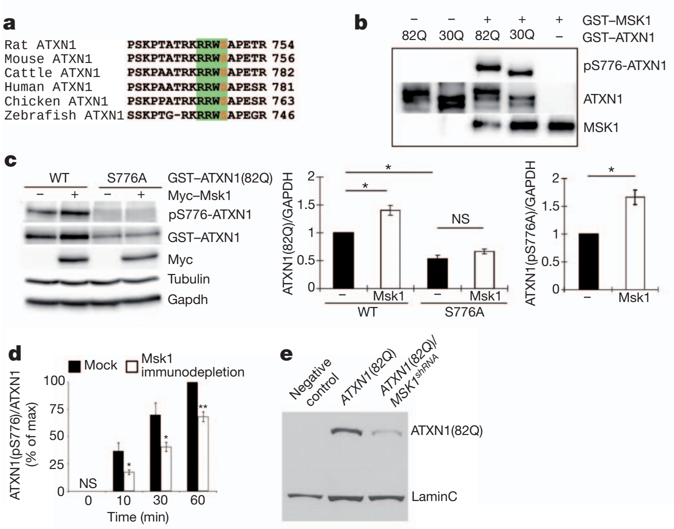


Figure 4 | MSK1 phosphorylates ATXN1 at S776 and controls its stability. **a**, MSK1 phosphorylation consensus site, RXXS, found in ATXN1 across different species. **b**, *In vitro* kinase assay with purified GST–ATXN1(82Q) or GST–ATXN1(30Q) and GST–MSK1. **c**, Neuro2a western blot transfected with Myc-fused mouse Msk1 and GST–ATXN1(82Q) or ATXN1(82Q, S776A). Graphs (right) show signal quantification of the western blot (error bars denote s.e.m. from three independent experiments, $*P < 0.05$). WT, wild type. **d**, Immunodepletion of Msk1 from cerebellar extracts decreases the level of phospho-S776 ATXN1. $*P < 0.05$, $**P < 0.01$, Student's *t*-test. **e**, Western blot shows that knockdown of the *Drosophila* homologue of *MSK1* decreases ATXN1(82Q) level in fly heads of indicated genotypes. LaminC, loading control.

In summary, the data from the cell-based screen, biochemical studies and genetic interactions strongly suggest that MSK1 phosphorylates ATXN1 and regulates its stability. Furthermore, we tested whether the components upstream of MSK1 in the MAPK pathway also affect ATXN1 S776 phosphorylation. Western blot analysis showed that knockdown of *RAS*, *FNTA* and *MEK* significantly decreases S776 phosphorylation, whereas constitutive activation of RAS and MEK1 markedly increases its phosphorylation and stability (Supplementary Fig. 4b, c and Supplementary Fig. 5).

Together, these results suggest that the RAS–MAPK–MSK1 pathway regulates ATXN1 stability through S776 phosphorylation.

Inhibition of MAPK pathway decreases ATXN1

Our data indicate that genetic inhibition of the MAPK pathway reduces ATXN1 levels and phenotypes. To evaluate the MAPK pathway as a potential therapeutic target for SCA1, we investigated whether pharmacological inhibition of the MAPK pathway (Fig. 3f) decreases ATXN1 levels. Treating human cells stably expressing ATXN1(82Q) with PD184352 (a MEK1/2 inhibitor[31]) or GW5704 (a RAF1 inhibitor[32]) decreased ATXN1(82Q) levels (Fig. 5a). The MSK1 inhibitors H89 and Ro-31-8220[33] also decreased the levels of ATXN1(82Q) (Fig. 5a).

To examine this reduction in a more physiologically relevant system, we tested whether inhibiting MAPK or MSK1 signalling affected Atxn1(82Q) abundance in the context of cerebellar slice cultures from SCA1 (82Q) knock-in mice[34]. Western blot analysis (Fig. 5b, c) showed that inhibition of Mek1/2 or Msk1 (by addition of PD184352 and Ro-31-8220, respectively) in cerebellar slice cultures decreased the levels of mutant Atxn1.

Together with the genetic data, these results strongly support the hypothesis that the MAPK/MSK1 pathway regulates ATXN1 levels. These findings also validate the genetic screen strategy and provide evidence that ATXN1 levels can be modulated pharmacologically.

Reducing MSK1 rescues degeneration in SCA1 mice

We next sought to test the genetic interaction between MSK1 and ATXN1 in mice. For this we used *Atxn1(154Q)* knock-in mice ($Atxn1^{154Q/+}$), which bear 154 CAG repeats in the endogenous mouse locus and recapitulate many features of human SCA1 (ref. 35). These mice were bred to *Msk1*$^{+/-}$ *Msk2*$^{+/-}$ mice, which show no obvious phenotypes[36,37]. First we analysed whether *Msk1* knockout would decrease Atxn1(154Q) in *Atxn1*$^{154Q/+}$ mice. Indeed, *Atxn1*$^{154Q/+}$ *Msk1*$^{-/-}$ animals presented lower levels of expanded Atxn1 than *Atxn1*$^{154Q/+}$ animals (Fig. 6a, b).

Atxn1$^{154Q/+}$ mice develop a motor phenotype that can be quantified after 9 weeks of age using a rotarod test. *Drosophila* has only one *Msk* gene (*JIL-1*), but mice have two paralogues: *Msk1* and *Msk2*. To account for potential genetic redundancy, we investigated the effect on the Atxn1(154Q)-induced phenotype of decreasing both genes simultaneously. The *Atxn1*$^{154Q/+}$ *Msk1*$^{+/-}$ *Msk2*$^{+/-}$ animals showed significantly better rotarod performance than their *Atxn1*$^{154Q/+}$ littermates ($P = 0.027$, Fig. 6c).

Next, we assessed the potential effect of decreasing Msk on Purkinje cell degeneration. Because Purkinje cell loss takes over 7–8 months to manifest in the *Atxn1*$^{154Q/+}$ mice, we resorted to another SCA1 mouse model that expresses ATXN1(82Q) using the *Pcp2* promoter (*B05* transgenic mice) and develops severe Purkinje cell degeneration after 12 weeks[38]. We found that eliminating one copy of mouse *Msk1* partially suppressed this Purkinje-cell-loss phenotype. Simultaneous elimination of one copy of *Msk1* and *Msk2* (*B05*$^{/+}$ *Msk1*$^{+/-}$ *Msk2*$^{+/-}$) also suppressed Purkinje cell loss (Fig. 6d). Moreover, *B05*$^{/+}$ *Msk1*$^{+/-}$ and *B05*$^{/+}$ *Msk1*$^{+/-}$ *Msk2*$^{+/-}$ mice have improved calbindin immunoreactivity in Purkinje cell dendrites compared to *B05*$^{/+}$ ($P < 0.001$ and $P < 0.01$, respectively). Notably, *B05*$^{/+}$ *Msk1*$^{+/-}$ *Msk2*$^{+/-}$ animals showed decreased levels of ATXN1(82Q) compared to *B05*$^{/+}$ controls (Supplementary Fig. 8b).

Discussion

Here we combine cross-species genetic screens and validation in human, mouse and *Drosophila* systems to identify previously unappreciated modulators of the steady-state levels of a neurodegeneration-causing protein. Two key discoveries highlight the power and reliability of this screening strategy. First, several modulators of ATXN1 levels belong to a single pathway, the RAS–MAPK–MSK1 signalling pathway. The functional relationship between the newly identified ATXN1 modulators underscores the validity of the approach and distinguishes these findings from screens done in one system whereby the large number of hits often precludes identifying the true positives and linking them in a functional pathway. Second, the screening approach and subsequent characterization revealed a direct biochemical mechanism governing ATXN1 levels (ATXN1 S776 phosphorylation by MSK1) as well as upstream regulators of ATXN1 abundance. The identification of direct (MSK1) and upstream (RAS and MAPK) regulators of ATXN1 levels provides several entry points for developing strategies for therapeutic intervention for SCA1.

Identifying multiple targets that can decrease disease protein levels is important for several reasons. First, the discovery of multiple therapeutic entry points counterbalances the frequent attrition of therapeutic targets during drug development. Second, the discovery of multiple targets enables the development of combination strategies (via multiple inhibitors or inhibitors with activity towards multiple targets[39]) to balance symptom mitigation and side effects. We note that, in the case of ATXN1, S776 can be phosphorylated by protein kinase A[29]. The discovery that MSK1 also phosphorylates ATXN1 at S776 raises the possibility that mild-to-modest inhibition of both kinases might be more effective than severe inhibition of either. Of course, the availability of several additional modulators of ATXN1 levels in the RAS–MAPK–MSK1 pathway now provides further opportunity to explore the development of combination therapeutics.

Previous efforts aimed at developing unbiased screens for the discovery of therapeutic targets for neurodegenerative diseases have focused on protein aggregation or suppression of neuronal cell death[40]. The strategy we used in this study focuses on targeting an earlier event in disease pathogenesis: decreasing the levels of the disease-causing protein. Genetic data in humans and animal models provide evidence that increased levels of either the mutant or the normal proteins are at the root of the pathogenesis of several neurodegenerative diseases. For example, expanded polyglutamine tracts stabilize the disease proteins in Huntington's disease and spinocerebellar ataxias[41–43], whereas increased levels of wild-type APP and α-synuclein due to genomic duplications and triplications cause Alzheimer's disease and Parkinson's disease, respectively[9–11]. Therefore, targeting an early event in pathogenesis could potentially delay the disease onset for this class of

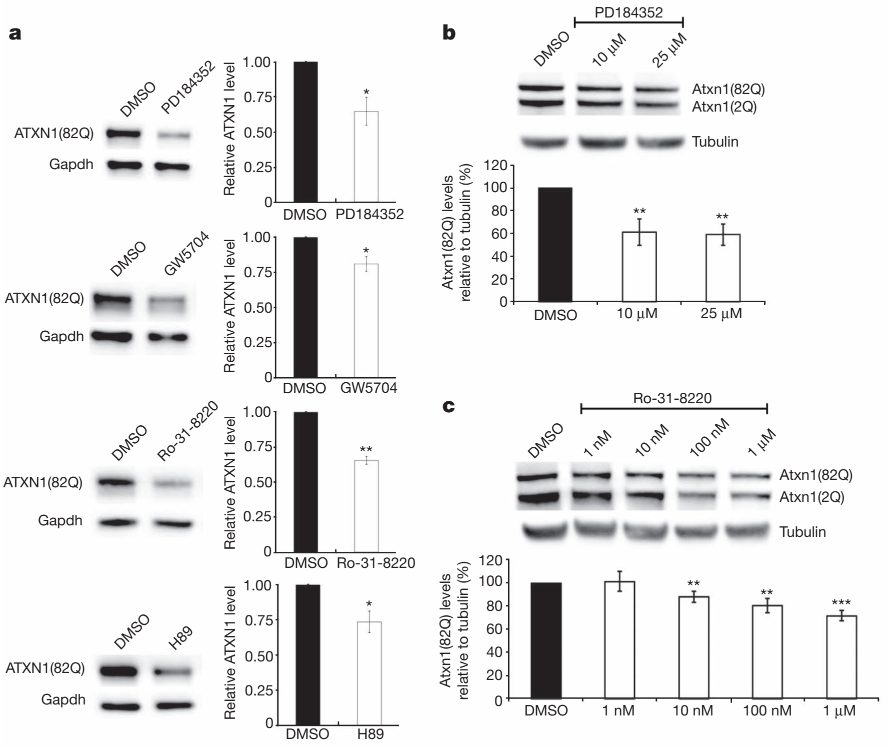


Figure 5 | Pharmacological inhibition of the MAPK pathway decreases ATXN1 level. a, Western blots and quantifications show a decrease in ATXN1 levels from Daoy mRFP–ATXN1(82Q) cells upon treatment with the indicated inhibitors (PD184352, 10 μM; GW5704, 10 μM; Ro-31-8220, 1 μM; H89, 5 μM). Error bars indicate s.e.m. from three independent experiments. **b**, **c**, Dose–response graphs show a decrease in Atxn1 levels from mouse cerebellar slices upon treatment with the indicated inhibitors. Error bars indicate s.e.m. from three independent experiments. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Student's *t*-test. DMSO, dimethylsulphoxide.

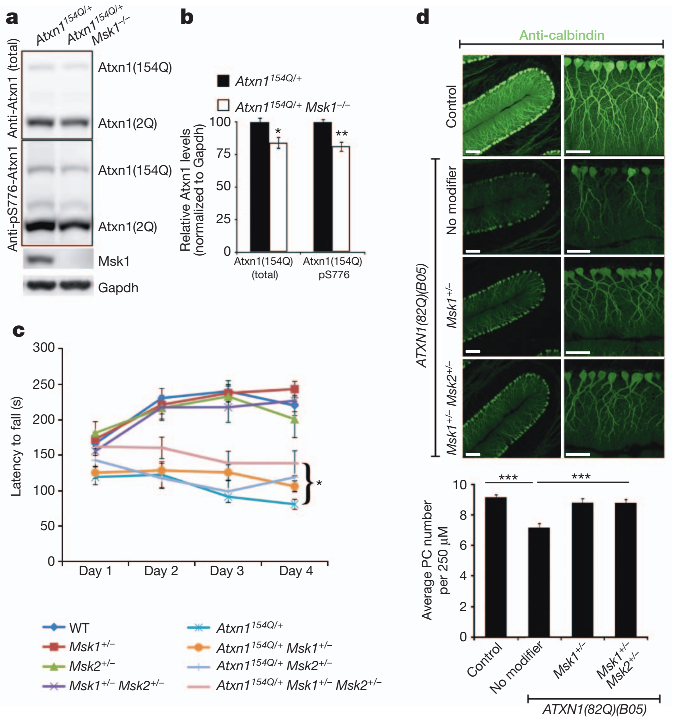


Figure 6 | Msk reduction rescues behavioural and pathological phenotypes in SCA1 mice. a, b, Western blot and quantification show decreased Atxn1 levels in *Atxn1*$^{154Q/+}$ cerebella upon complete loss of Msk1 (4–5-week-old mice; error bars denote s.e.m. from $n = 4$ per genotype). **c**, Reduction of Msk1 and Msk2 improves motor performance of *Atxn1*$^{154Q/+}$ (9–10 weeks old, see Supplementary Information for details, $*P = 0.027$). **d**, Partial loss-of-function of Msk1 alone or Msk1 and Msk2 rescues Purkinje cell (PC) loss of *ATXN1(82Q) (B05)* mice (12 weeks old; $n = 3$ per genotype, three sections per mice). Scale bars, 100 μm (left), 50 μm (right). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, post-hoc Holm-Šídák test.

disorders. Indeed, a key challenge in Alzheimer's and Parkinson's diseases has been our limited understanding of how the stability of APP and α-synuclein is controlled. An unbiased forward genetic screen using independent and complementary assays to decrease the levels of APP and α-synuclein is likely to yield a variety of targets that might change the landscape of therapeutic development for these diseases. Furthermore, applying this approach at a genome-wide scale would enable us to discover other possible therapeutic targets. We note that cross-species comparative genomics (messenger RNA expression profiling and gene copy number) has been recently used to discover new pathogenic pathways in the cancer field[44]. We believe that this study provides a proof-of-principle for addressing other neurodegenerative diseases using a similar conceptual strategy.

METHODS SUMMARY

The Daoy mRFP–ATXN1(82Q)–IRES–YFP stable cell line was generated by cloning into pHAGE lentiviral vector, then selected with puromycin and run through AriaII (BD Biosciences) for selection of single cells that showed expression of mRFP and YFP. For the kinase siRNA screen, cells were split into 96-well plates and each siRNA (kinase siRNA library from Invitrogen) was transfected into corresponding wells. After 72 h incubation, the cells were analysed by flow cytometry (LSR II, BD Biosciences). Statistical analyses for primary and secondary human cell screens were performed using JMP software from SAS. For the *Drosophila* genetic screen, we used the *GMR* > *ATXN1(82Q)* fly line[22]. In total, 704 kinase RNA interference or loss-of-function alleles were obtained from VDRC, Bloomington and Harvard repositories. Scanning electron microscopy and paraffin sections of the retina were performed as previously described[22]. All *Drosophila* genotypes are indicated in Supplementary Information. The *Drosophila* climbing assay was done as previously described[28]. For kinase assay, 1 μg of bacterially expressed and purified GST–ATXN1(82Q) or GST–ATXN1(30Q) and 100 ng of active MSK1 (Invitrogen) were incubated in kinase reaction buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM dithiothreitol) with phosphatase inhibitor (Roche) and 30 μM ATP (Invitrogen) for 30 min at 30 °C. For mice cerebella sections, cerebellum from SCA1 (82Q) knock-in mice[34] at post-natal day 11 were embedded in a 2% agarose/Gey's balanced salt solution containing 1 mM kynurenic acid (GBSSK) solution and sectioned at 350 μm as described[45]. Slices were cultured in collagen coated membrane inserts in cerebellar slice culture media. *Atxn1*$^{154Q/+35}$ or *ATXN1(82Q)* (B05)[38] and *Msk1*$^{+/-}$ *Msk2*$^{+/-36,37}$ mice were crossed to obtain mice with appropriate genotypes. Rotarod analysis was performed as previously described[46]. Purkinje cell pathology was observed by immunofluorescence staining of cerebellar sections using anti-calbindin antibody (Sigma) and imaged by a Zeiss LSM 710 confocal microscope[47].

Full Methods and any associated references are available in the online version of the paper.

Received 6 August 2012; accepted 19 April 2013.
Published online 29 May 2013.

1. Evans, D. A. Estimated prevalence of Alzheimer's disease in the United States. *Milbank Q.* **68,** 267–289 (1990).
2. Hindle, J. V. Ageing, neurodegeneration and Parkinson's disease. *Age Ageing* **39,** 156–161 (2010).
3. Marsden, C. D. & Parkes, J. D. Success and problems of long-term levodopa therapy in Parkinson's disease. *Lancet* **1,** 345–349 (1977).
4. Scarpini, E., Scheltens, P. & Feldman, H. Treatment of Alzheimer's disease: current status and new perspectives. *Lancet Neurol.* **2,** 539–547 (2003).
5. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nature Med.* **10,** S10–S17 (2004).
6. Taylor, J. P., Hardy, J. & Fischbeck, K. H. Toxic proteins in neurodegenerative disease. *Science* **296,** 1991–1995 (2002).
7. Haass, C. & Selkoe, D. J. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nature Rev. Mol. Cell Biol.* **8,** 101–112 (2007).
8. Zoghbi, H. Y. & Orr, H. T. Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.* **23,** 217–247 (2000).
9. Singleton, A. B. *et al.* α-synuclein locus triplication causes Parkinson's disease. *Science* **302,** 841 (2003).
10. Chartier-Harlin, M. C. *et al.* α-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364,** 1167–1169 (2004).
11. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genet.* **38,** 24–26 (2006).
12. Götz, J. & Ittner, L. M. Animal models of Alzheimer's disease and frontotemporal dementia. *Nature Rev. Neurosci.* **9,** 532–544 (2008).
13. Williams, A. J. & Paulson, H. L. Polyglutamine neurodegeneration: protein misfolding revisited. *Trends Neurosci.* **31,** 521–528 (2008).
14. Xia, H. *et al.* RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature Med.* **10,** 816–820 (2004).
15. Harper, S. Q. *et al.* RNA interference improves motor and neuropathological abnormalities in a Huntington's disease mouse model. *Proc. Natl Acad. Sci. USA* **102,** 5820–5825 (2005).
16. Yamamoto, A., Lucas, J. J. & Hen, R. Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell* **101,** 57–66 (2000).
17. Zu, T. *et al.* Recovery from polyglutamine-induced neurodegeneration in conditional *SCA1* transgenic mice. *J. Neurosci.* **24,** 8853–8861 (2004).
18. Kordasiewicz, H. B. *et al.* Sustained therapeutic reversal of Huntington's disease by transient repression of huntingtin synthesis. *Neuron* **74,** 1031–1044 (2012).
19. Zoghbi, H. Y. & Orr, H. T. Pathogenic mechanisms of a polyglutamine-mediated neurodegenerative disease, spinocerebellar ataxia type 1. *J. Biol. Chem.* **284,** 7425–7429 (2009).
20. Emamian, E. S. *et al.* Serine 776 of ataxin-1 is critical for polyglutamine-induced disease in *SCA1* transgenic mice. *Neuron* **38,** 375–387 (2003).
21. Noble, M. E., Endicott, J. A. & Johnson, L. N. Protein kinase inhibitors: insights into drug design from structure. *Science* **303,** 1800–1805 (2004).
22. Fernandez-Funez, P. *et al.* Identification of genes that modify ataxin-1-induced neurodegeneration. *Nature* **408,** 101–106 (2000).
23. Taniguchi, C. M., Emanuelli, B. & Kahn, C. R. Critical nodes in signalling pathways: insights into insulin action. *Nature Rev. Mol. Cell Biol.* **7,** 85–96 (2006).
24. Shaharabany, M. *et al.* Distinct pathways for the involvement of WNK4 in the signaling of hypertonicity and EGF. *FEBS J.* **275,** 1631–1642 (2008).
25. Teixeira-Castro, A. *et al.* Neuron-specific proteotoxicity of mutant ataxin-3 in *C. elegans*: rescue by the DAF-16 and HSF-1 pathways. *Hum. Mol. Genet.* **20,** 2996–3009 (2011).
26. Cobb, M. H. MAP kinase pathways. *Prog. Biophys. Mol. Biol.* **71,** 479–500 (1999).
27. Avruch, J. Insulin signal transduction through protein kinase cascades. *Mol. Cell. Biochem.* **182,** 31–48 (1998).
28. Al-Ramahi, I. *et al.* dAtaxin-2 mediates expanded Ataxin-1-induced neurodegeneration in a *Drosophila* model of SCA1. *PLoS Genet.* **3,** e234 (2007).
29. Jorgensen, N. D. *et al.* Phosphorylation of ATXN1 at Ser776 in the cerebellum. *J. Neurochem.* **110,** 675–686 (2009).
30. Roux, P. P. & Blenis, J. ERK and p38 MAPK-activated protein kinases: a family of protein kinases with diverse biological functions. *Microbiol. Mol. Biol. Rev.* **68,** 320–344 (2004).
31. Mody, N., Leitch, J., Armstrong, C., Dixon, J. & Cohen, P. Effects of MAP kinase cascade inhibitors on the MKK5/ERK5 pathway. *FEBS Lett.* **502,** 21–24 (2001).

32. Lackey, K. *et al.* The discovery of potent cRaf1 kinase inhibitors. *Bioorg. Med. Chem. Lett.* **10,** 223–226 (2000).
33. Deak, M., Clifton, A. D., Lucocq, L. M. & Alessi, D. R. Mitogen- and stress-activated protein kinase-1 (MSK1) is directly activated by MAPK and SAPK2/p38, and may mediate activation of CREB. *EMBO J.* **17,** 4426–4441 (1998).
34. Lorenzetti, D. *et al.* Repeat instability and motor incoordination in mice with a targeted expanded CAG repeat in the *Sca1* locus. *Hum. Mol. Genet.* **9,** 779–785 (2000).
35. Watase, K. *et al.* A long CAG repeat in the mouse *Sca1* locus replicates SCA1 features and reveals the impact of protein solubility on selective neurodegeneration. *Neuron* **34,** 905–919 (2002).
36. Arthur, J. S. & Cohen, P. MSK1 is required for CREB phosphorylation in response to mitogens in mouse embryonic stem cells. *FEBS Lett.* **482,** 44–48 (2000).
37. Wiggin, G. R. *et al.* MSK1 and MSK2 are required for the mitogen- and stress-induced phosphorylation of CREB and ATF1 in fibroblasts. *Mol. Cell. Biol.* **22,** 2871–2881 (2002).
38. Burright, E. N. *et al. SCA1* transgenic mice: a model for neurodegeneration caused by an expanded CAG trinucleotide repeat. *Cell* **82,** 937–948 (1995).
39. Dar, A. C., Das, T. K., Shokat, K. M. & Cagan, R. L. Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature* **486,** 80–84 (2012).
40. van Ham, T. J., Breitling, R., Swertz, M. A. & Nollen, E. A. Neurodegenerative diseases: lessons from genome-wide screens in small model organisms. *EMBO Mol. Med.* **1,** 360–370 (2009).
41. Jeong, H. *et al.* Acetylation targets mutant huntingtin to autophagosomes for degradation. *Cell* **137,** 60–72 (2009).
42. Matsumoto, M. *et al.* Molecular clearance of ataxin-3 is regulated by a mammalian E4. *EMBO J.* **23,** 659–669 (2004).
43. Cummings, C. J. *et al.* Mutation of the E6-AP ubiquitin ligase reduces nuclear inclusion frequency while accelerating polyglutamine-induced pathology in *SCA1* mice. *Neuron* **24,** 879–892 (1999).
44. Chin, L. & Gray, J. W. Translating insights from the cancer genome into clinical practice. *Nature* **452,** 553–563 (2008).
45. Falsig, J. & Aguzzi, A. The prion organotypic slice culture assay—POSCA. *Nature Protocols* **3,** 555–562 (2008).
46. Jafar-Nejad, P., Ward, C. S., Richman, R., Orr, H. T. & Zoghbi, H. Y. Regional rescue of spinocerebellar ataxia type 1 phenotypes by 14–3-3 haploinsufficiency in mice underscores complex pathogenicity in neurodegeneration. *Proc. Natl Acad. Sci. USA* **108,** 2142–2147 (2011).
47. Bowman, A. B. *et al.* Duplication of *Atxn1l* suppresses SCA1 neuropathology by decreasing incorporation of polyglutamine-expanded ataxin-1 into native complexes. *Nature Genet.* **39,** 373–379 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the members of the Zoghbi, Botas, Orr and Westbrook laboratories for suggestions and discussions, and V. Brandt for editorial input. We also appreciate the help from CCSC and C-BASS cores at The Baylor College of Medicine (BCM) for FACS analysis and the confocal microscopy and mouse behavioural cores of the BCM Intellectual and Developmental Disabilities Research Center (HD024064). Thanks to J. Barrish at Texas Children's Hospital for help with scanning electron microscopy. This work was supported by a Howard Hughes Medical Institute Collaborative Innovation Awards grant and grant NIH-NS42179 (J.B.) I.A.-R. was supported by an NIH Brain Disorders and Development training grant.

Author Contributions J.P., I.A.-R., Q.T., N.M., H.T.O., T.F.W., J.B. and H.Y.Z. designed the experiments. J.P., I.A.-R., Q.T., N.M., J.R.D.-G., T.G.-F., H.-C.L., S.L., L.D., H.K., Y.L., P.J.-N., L.S.S., R.R., X.L. and Y.G. performed the research. J.P., I.A.-R., Q.T., N.M., J.R.D.-G., T.G.-F., H.-C.L., S.L., L.D., H.K., Y.L., P.J.-N., C.A.S., J.S.C.A., H.T.O., T.F.W., J.B. and H.Y.Z. analysed and interpreted the data. J.P., I.A.-R., H.T.O., T.F.W., J.B. and H.Y.Z. wrote and edited the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.T.O. (orrxx002@umn.edu), T.F.W. (thomasw@bcm.edu), J.B. (jbotas@bcm.edu) or H.Y.Z. (hzoghbi@bcm.edu).

METHODS

Generation of stable cell lines. mRFP–ATXN1(82Q)–IRES–YFP was cloned into pHAGE vector. Lentiviral packaged clones were infected into Daoy cells and then were selected with puromycin and ran through Aria II (BD Biosciences) for selection of cells that shows expressions of mRFP and YFP. Other cell clones were generated by the same method: Daoy mRFP–ATXN1(30Q)–IRES–YFP and Daoy mRFP–IRES–YFP.

Cell-based kinase siRNA screen. Daoy mRFP–ATXN1(82Q)–IRES–YFP was split into 96-well plates. On the next day, each siRNA (kinase siRNA library from Invitrogen) was transfected at 20 nM with 0.08 μl of transfection reagent (Dharmacon) into corresponding wells and incubated for 72 h. Before running FACS analysis (LSR II, BD Biosciences), the cells were trypsinized and suspended in PBS with 5% FBS.

***Drosophila* kinase screen.** For the screen we used the previously characterized *y,w,UAS-ATXN1(82Q)*(line-F7)*;GMR-GAL4* line[22] and for testing polyglutamine specificity we used *y,w,UAS-ATXN1(30Q)* (line-F1)*;GMR-GAL4* (ref. 22). RNAi or loss-of-function alleles were obtained from the Vienna *Drosophila* RNAi Center (http://stockcenter.vdrc.at/control/main) and the Bloomington *Drosophila* Stock Center at University of Indiana (http://flystocks.bio.indiana.edu) repositories. Animals were crossed at 28 °C for external eye phenotype and 25 °C for retinal experiments. Scanning electron microscopy and paraffin sections of the retina were performed as previously described[22].

Cell culture and siRNA transfections. Daoy stable cell lines and Neuro2a cell lines were cultured in DMEM (Invitrogen) with 10% FBS (Invitrogen). siRNAs (Invitrogen) were transfected with DharmaFECT (Dharmacon) and incubated for 3 days before analysis. GST-fused human ATXN1(82Q) and ATXN1(82Q, S776A)[48], and Myc-fused mouse Msk1 (Open Biosystems), MEK-DD (gift of W. Hahn, Addgene plasmid no. 15268) and HRAS-V12 (gift of W. Hahn, Addgene plasmid no. 9051) were transfected by Lipofectamine 2000 (Invitrogen). Inhibitors used: PD184352 (Santa Cruz Biotechnology), GW5074 (Sigma), H89 (Sigma) and Ro-31-8220 (Calbiochem).

Cell lysate preparation and immunoblot analysis. Before collection, cells were washed with PBS and lysed on ice for 20 min in radioimmunoprecipitation assay (RIPA) buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% NP-40) supplemented with protease inhibitors (Roche). The cell lysates were centrifuged at 15,000*g* for 20 min at 4 °C, and the supernatants were analysed by SDS–PAGE and western blot. Primary antibodies used: anti-ATXN1 (11750), anti-pS776 ATXN1 (PN1248)[20], anti-GFP (Genetex), anti-Myc (Sigma), anti-GST (Sigma), anti-Gapdh (Advanced ImmunoChemical), anti-MEK (BD Biosciences), anti-Ras (BD Biosciences), anti-tubulin (Abcam) and anti-MSK1 (R&D Systems).

Kinase assay. Complementary DNAs of *ATXN1(82Q)* and *ATXN1(30Q)* were cloned into pDEST15 vector (GST-tagged, Invitrogen), and then transformed into BL21. GST–ATXN1(82Q) and GST–ATXN1(30Q) were purified through a GST column. One microgram of GST–ATXN1(82Q) or (30Q) and 100 ng of active MSK1 (Invitrogen), RSK1 (Sigma) or RSK2 (Sigma) were incubated in kinase reaction buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM dithiothreitol (DTT)) with phosphatase inhibitor (Roche) and 30 μM ATP (Invitrogen) for 30 min at 30 °C. The kinase reaction was terminated by adding both NuPAGE LDS sample buffer (Invitrogen) and sample reducing agent (Invitrogen) and then boiled for 10 min. The samples were run immediately onto NuPAGE Novex 4–12% Bis-Tris Gel (Invitrogen) for western blot analysis. Antibodies used were anti-ATXN1 (11750), anti-pS776 ATXN1 (PN1248)[20], anti-pan RSK1/2/3 (Cell Signaling Technology) and anti-MSK1 (R&D Systems).

***Drosophila* motor performance tests.** Motor performance tests were done as previously described[28] with some modifications. Fifteen age-matched virgin females were placed in a vial and tapped down. The number of flies that climbed up 9 cm in 15 s was recorded. We repeated this ten consecutive times and the average of the ten observations was plotted for each day as shown in the chart. Two replicates were tested in parallel for each genotype. ATXN1(82Q) was expressed in the nervous system using *nrv2*-Gal4 (*y,w,UAS-ATXN1(82Q)* (line-F7); *nrv2-GAL4*).

Preparation of *Drosophila* protein lysates and immunoblot. Protein lysates were prepared in NuPAGE LDS sample buffer (Invitrogen), immunoblotted in acrylamide gels following standard procedures. Primary antibodies used are 11NQ anti-ATXN1 (Neuromab) and anti-laminC (Developmental Studies Hybridoma Bank).

Mouse brain slice culture. Mouse cerebella at postnatal day 11 were embedded in a 2% agarose GBSSK solution and sectioned at 350 μm as described[45]. Slices were cultured in collagen coated membrane inserts in cerebellar slice culture media. Half media changes were performed every 2–3 days, and PD184352 and Ro-31-8220 drugs were replaced accordingly.

Immunodepletion study. Dissected cerebella from FVB mice were homogenized in 100 μl lysis buffer (50 mM Tris, pH 7.5, 100 mM NaCl, 2.5 mM $MgCl_2$, 0.5% Triton X-100) containing protease inhibitor cocktail (Roche Biochemicals) and phosphatase inhibitor cocktails II and III (Sigma). Homogenates were frozen and thawed three times and cleared at 15,000 r.p.m. for 10 min. For the immunodepletion, 150 μg of cleared cerebellar lysate was combined with 10 μl of PBS or goat anti-MSK1 antibody (R&D Systems, AF2518) and lysis buffer in a total volume of 100 μl for a minimum of 24 h at 4 °C with rotation. Fifty microlitres of a 50% slurry of Protein G Sepharose beads (Sigma) were added to the samples for a minimum of 48 h with regular rotation at 4 °C. Complete immunodepletion of the lysates was assessed by western blotting of the cleared lysates and proteins bound to the beads by boiling the beads in reducing sample buffer. Ten micrograms of mock- or kinase-depleted samples were combined with 10 mM $MgCl_2$ and 300 ng of GST-tagged ATXN1(30Q) protein; kinase reactions were initiated with the addition of 200 mM ATP followed by incubation at 30 °C for 0, 10, 30 or 60 min. Activity was stopped by boiling in Laemmli sample buffer for 10 min. Phosphorylation was quantified by western blot as previously described[29]. Data from five kinase assays from each set of immunodepletions were plotted, and statistical significance was determined using a Student's *t*-test.

Statistical analyses. To identify the primary screen hits, we calculated the whole-screen mean and selected the siRNAs that decreased the mRFP–ATXN1(82Q)/YFP ratio below 2 standard deviations. For the confirmation screen, we performed three independent experimental sets and compared the effect of siRNAs within each set to the internal siRNA controls by analysis of variance followed by Dunnet's and Tukey's post-hoc tests to select for the siRNAs that were significantly decreased. The same method was used to identify the false-positive siRNAs, by testing their effect on the control reporter mRFP–IRES–YFP cell line. All statistical analyses were performed using the JMP software from SAS.

Quantitative RT–PCR. RNA from siRNA-transfected Daoy (82Q) stable cells was extracted using Trizol (Invitrogen). Random-primed cDNA was obtained using the Superscript III kit (Invitrogen). Quantitative RT–PCR was performed using Perfecta SYBR Green FastMix (Quanta Biosciences). Primers used were as follows: *FNTA*: forward primer 5′-TGGACGACGGGTTTGTGAG-3′, reverse primer 5′-ACCGGATCTATATCAGCCCATT-3′; *MEK2*: forward 5′-CCAAGGTCGGCGAACTCAAA-3′, reverse 5′-TCTCAAGGTGGATCAGCTTCC-3′; *MEK3*: forward 5′-GTCGACTGTTTCTACACTGTC-3′, reverse 5′-GGATGTCCTCTGGAATTGTC-3′. siRNAs (Invitrogen) used were as follows: *siIGF1R* (1): sense 5′-UCUUCAAGGGCAAUUUGCUCAUUAA-3′, antisense 5′-UUAAUGAGCAAAUUGCCCUUGAAGA-3′; *siIGF1R* (2): sense 5′-CAACACUGGUCAUCAUGGAACUGAU-3′, antisense 5′-AUCAGUUCCAUGAUGACCAGUGUUG-3′; *siULK3* (1): sense 5′-GGGACAGUGACAAUAUCUACCUCAU-3′, antisense 5′-AUGAGGUAGAUAUUGUCACUGUCCC-3′; *siULK3* (2): sense 5′-UCGCUUCAUCCAUACCCGCAGGAUU-3′, antisense 5′-AAUCCUGCGGGUAUGGAUGAAGCGA-3′; *siWNK4* (1): sense 5′-UGGGCUUGGUCUGUGAAGCCGAUUA-3′, antisense 5′-UAAUCGGCUUCACAGACCAAGCCCA-3′; *siWNK4* (2): sense 5′-GCGAAAGCGUGAGAAGCUGCGUAAA-3′, antisense 5′-UUUACGCAGCUUCUCACGCUUUCGC-3′; *siBUB1*: sense 5′-GCUGCACAACUUGCGUCUACACCAU-3′, antisense 5′-AUGGUGUAGACGCAAGUUGUGCAGC-3′; *siMSK1* (1): sense 5′-UCCUUUGUUGCUCCUUCCAUCCUAU-3′, antisense 5′-AUAGGAUGGAAGGAGCAACAAAGGA-3′; *siMSK1* (2): sense 5′-CCUUUGUUGCUCCUUCCAUCCUAUU-3′, antisense 5′-AAUAGGAUGGAAGGAGCAACAAAGG-3′; *siMEK2*: sense 5′-CGACUUCCAGGAGUUUGUCAAUAAA-3′, antisense 5′-UUUAUUGACAAACUCCUGGAAGUCG-3′; *siMEK3*: sense 5′-CCUUCAUCACCAUUGGAGACAGAAA-3′, antisense 5′-UUUCUGUCUCCAAUGGUGAUGAAGG-3′; *siMEK6*: sense 5′-CGGCUACUGAUGGAUUUGGAUAUUU-3′, antisense 5′-AAAUAUCCAAAUCCAUCAGUAGCCG-3′; *siERK1*: 5′-CCCUGGAAGCCAUGAGAGAUGUCUA-3′, antisense 5′-UAGACAUCUCUCAUGGCUUCCAGGG-3′; *siERK2* (1): sense 5′-GCCGAAGCACCAUUCAAGUUCGACA-3′, antisense 5′-UGUCGAACUUGAAUGGUGCUUCGGC-3′; *siERK2* (2): sense 5′-CCGAAGCACCAUUCAAGUUCGACAU-3′, antisense 5′-AUGUCGAACUUGAAUGGUGCUUCGG-3′; *siHRAS* (1): sense 5′-CCAUCCAGCUGAUCCAGAACCAUUU-3′, antisense 5′-AAAUGGUUCUGGAUCAGCUGGAUGG-3′; *siHRAS* (2): sense 5′-GAUCAAACGGGUGAAGGACUCGGAU-3′, antisense 5′-AUCCGAGUCCUUCACCCGUUUGAUC-3′; *siFNTA* (1): sense 5′-GCAGGAUCGUGGUCUUUCCAAAUAU-3′, antisense 5′-AUAUUUGGAAAGACCACGAUCCUGC-3′; *siFNTA* (2): sense 5′-GACAAUGGGUUAUUCAGGAAUUUAA-3′, antisense 5′- UUAAAUUCCUGAAUAACCCAUUGUC; *siFNTA* (3): sense 5′-CAUAAUGAAAGUGCAUGGAACUAUU-3′, antisense 5′-AAUAGUUCCAUGCACUUUCAUUAUG-3′.

Mouse models. All procedures for mouse animal use were approved by the Institutional Animal Care and Use Committee for Baylor College of Medicine and Affiliates. *Atxn1*$^{154Q/+}$ has been previously described[35] and has been backcrossed to C57BL/6 for more than ten generations. FVB *ATXN1(82Q) B05* transgenic[38] and C57BL/6 $Msk1^{-/-}$ $Msk2^{-/-}$[36,37] mice have been previously generated and characterized. Male $Atxn1^{154Q/+}$ animals were crossed with female $Msk1^{+/-}$ $Msk2^{+/-}$ to obtain the following genotypes: wild type, $Msk1^{+/-}$, $Msk2^{+/-}$, $Msk1^{+/-}$ $Msk2^{+/-}$, $Atxn1^{154Q/+}$, $Msk1^{+/-}$ $Atxn1^{154Q/+}$, $Msk2^{+/-}$ $Atxn1^{154Q/+}$ and $Msk1^{+/-}$ $Msk2^{+/-}$ $Atxn1^{154Q/+}$. Male *ATXN1(82Q)(B05)* was crossed with female $Msk1^{+/-}$ $Msk2^{+/-}$ for the following F_1 genotypes: wild type, $Msk1^{+/-}$, $Msk2^{+/-}$, $Msk1^{+/-}$ $Msk2^{+/-}$, *ATXN1(82Q)(B05)*, *ATXN1(82Q)(B05)* $Msk1^{+/-}$, *ATXN1(82Q)(B05)* $Msk2^{+/-}$ and *ATXN1(82Q)(B05)* $Msk1^{+/-}$ $Msk2^{+/-}$.

Rotarod analysis. Rotarod analysis was performed as previously described[46], with four trials for 4 days using 9–10-week-old male mice. Data were analysed using post-hoc test.

Mouse brain lysate preparation and immunoblot analysis. Mice cerebella were dissected and then lysed in RIPA buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% NP-40) supplemented with protease inhibitors (Roche). The protein lysate was then incubated on ice for 20 min and centrifuged at 13,000 r.p.m. for 20 min at 4 °C, and the supernatants were analysed by SDS–PAGE and western blot. Primary antibodies used were anti-ATXN1 (11750), anti-pS776 ATXN1 (PN1248)[20], anti-Gapdh (Advanced ImmunoChemical), anti-pERK (Cell signaling), anti-tubulin (Abcam) and anti-MSK1 (R&D Systems).

Purkinje cell pathology analysis. Mice were perfused with 4% paraformaldehyde in PBS and whole brains were dissected. Cerebellar sections were cut at a thickness of 30 μm and were immunostained with anti-calbindin antibody (Sigma) and imaged using a Zeiss LSM 710 confocal microscope. Quantification was performed using ImageJ software as previously described[47]. Data were analysed using post-hoc test.

***Drosophila* lifespan assays.** Two replicates of 30 female virgins each were aged at 26.5 °C. Vials were counted every day for the duration of the experiment to score dead animals. Data were represented using Kaplan–Meyer and significance was assessed through the Wilcoxon test.

Functional enrichment and network analysis of modifiers. Functional enrichment was assessed using KEGG/DAVID analysis database (http://david.abcc.ncifcrf.gov/). We carried out network analysis using Ingenuity Pathway Analysis and manual curation.

Gel-filtration chromatography and kinase assay. Protein lysates were prepared from two C57BL/6 wild-type mouse cerebella and gel-filtration chromatography was performed as described previously[46,47]. Atxn1 carboxy-terminal fragment (the last 269 amino acids of Atxn1) fused to maltose binding protein (MBP) was purified from bacteria[49]. Kinase assay reactions were set up by combining 1% of each fractionated cerebellar lysate with 100 ng MBP–Atxn1 C-terminal fragment in kinase buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) with phosphatase inhibitor (Roche) and 30 μM ATP (Invitrogen) and then were incubated for 30 min at 30 °C. The kinase reaction was terminated by adding NuPAGE LDS sample buffer (Invitrogen) and sample reducing agent (Invitrogen) and then boiled for 10 min. The samples were run immediately onto NuPAGE Novex 4–12% Bis-Tris Gel (Invitrogen) for western blot analysis. S776 phosphorylation was detected by using anti-pS776 ATXN1 antibody (PN1248)[20].

48. Lim, J. *et al.* Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1. *Nature* **452,** 713–718 (2008).
49. Servadio, A. *et al.* Expression analysis of the ataxin-1 protein in tissues from normal and spinocerebellar ataxia type 1 individuals. *Nature Genet.* **10,** 94–98 (1995).

doi:10.1038/nature12305

Structural mechanism of cytosolic DNA sensing by cGAS

Filiz Civril[1]*, Tobias Deimling[1]*, Carina C. de Oliveira Mann[1], Andrea Ablasser[2], Manuela Moldt[1], Gregor Witte[1], Veit Hornung[2] & Karl-Peter Hopfner[1,3]

Cytosolic DNA arising from intracellular bacterial or viral infections is a powerful pathogen-associated molecular pattern (PAMP) that leads to innate immune host defence by the production of type I interferon and inflammatory cytokines. Recognition of cytosolic DNA by the recently discovered cyclic-GMP-AMP (cGAMP) synthase (cGAS) induces the production of cGAMP to activate the stimulator of interferon genes (STING). Here we report the crystal structure of cGAS alone and in complex with DNA, ATP and GTP along with functional studies. Our results explain the broad DNA sensing specificity of cGAS, show how cGAS catalyses dinucleotide formation and indicate activation by a DNA-induced structural switch. cGAS possesses a remarkable structural similarity to the antiviral cytosolic double-stranded RNA sensor 2′-5′oligoadenylate synthase (OAS1), but contains a unique zinc thumb that recognizes B-form double-stranded DNA. Our results mechanistically unify dsRNA and dsDNA innate immune sensing by OAS1 and cGAS nucleotidyl transferases.

Recognition of pathogen- or danger-associated molecular patterns (PAMPs or DAMPs) is crucial for host defence. Innate immunity ensures this recognition through germline-encoded pattern recognition receptors (PRRs) and triggers signalling cascades that result in production of proinflammatory cytokines and type I interferons (IFN-α and IFN-β)[1,2]. Cytosolic DNA arising from intracellular bacteria or viral infections is a powerful PAMP and is also implicated as a DAMP in autoimmune diseases[1,3,4]. Over the past years, a variety of PRRs for cytosolic DNA have been reported: DNA-dependent activator of IFN-regulatory factors (DAI, also known as ZBP1)[5], absent in melanoma 2 (AIM2)[6–8], RNA polymerase III[9,10], leucine-rich repeat (in Flightless I) interacting protein-1 (LRRFIP1)[11], DExD/H box helicases (DDX41, DHX9 and DHX36)[12,13] and IFN-inducible protein IFI16[14]. However, these PRRs are either cell-type- or DNA-sequence-specific, are possible accessory factors (DExD/H proteins), or trigger different pathways such as caspase-1 activation (AIM2) or a β-catenin-dependent signalling pathway (LRRFIP1)[15].

Although the DNA sensor for type I IFN production with broad specificity and cell distribution was not identified until recently, it was known that IRF3 and NFκB activation in response to DNA requires STING (stimulator of interferon genes, encoded by gene *TMEM173* the protein is also known as MITA, MPYS or ERIS), a transmembrane protein that is resident on the endoplasmic reticulum[16–18]. STING colocalizes with DNA *in vivo* but binds DNA only with low affinity *in vitro*[19], suggesting the presence of an additional sensor. Furthermore, STING is a direct PRR for cyclic dinucleotides such as c-di-AMP and c-di-GMP[20], which are signalling molecules in prokaryotes and trigger IFN in response to, for example, intracellular bacteria[21,22].

Recent results identified human c-GMP-AMP (cGAMP) synthase (cGAS, also known as C6ORF150 and male abnormal 21 domain containing 1 (MB21D1)) as a broad-specificity cytosolic DNA sensor[23]. In the presence of DNA cGAS produces cGAMP, which is an endogenous second messenger that activates STING[18], explaining how STING can stimulate IFN in response to both cyclic dinucleotides and DNA. To reveal the mechanism of DNA-stimulated cGAMP synthesis, we determined the crystal structure of porcine $cGAS^{Mab21}$ (residues 135–497, comprising the highly conserved, DNA-stimulated nucleotidyl transferase (NTase) domain) with and without a 14-mer dsDNA ligand and nucleotide substrates, along with functional studies *in vitro* and in living cells.

Crystal structure of $cGAS^{Mab21}$

cGAS is a 60 kDa protein composed of an unstructured, not well conserved amino-terminal stretch of approximately 130–150 residues followed by a highly conserved Mab21 domain that belongs to the nucleotidyl transferase (NTase) superfamily[24]. To overproduce and crystallize cGAS, it was necessary to genetically remove the unstructured N-terminal tail. The resulting $cGAS^{Mab21}$ used in this study (residues 155/161–522 for human cGAS and residues 135–497 for porcine cGAS) possesses DNA-dependent dinucleotide synthesis activity in the presence of a 50-mer dsDNA that induces IFN in THP1 cells (Fig. 1a and Supplementary Fig. 1a, b). Whereas cGAS also produces cGAMP in the presence of a 40-mer dsDNA, no activity was observed when we omitted either GTP or ATP from the reaction mixture or substituted dsDNA with single-stranded DNA (Supplementary Fig. 1a).

We determined the crystal structure of porcine $cGAS^{Mab21}$ by single-wavelength anomalous dispersion to 2.5 Å resolution using a selenomethionine derivative. After density modification, we could build an initial model, which was completed and refined against the 2.0 Å resolution native data, resulting in good *R*-factors and stereochemistry (Supplementary Fig. 1c and Supplementary Table 1).

The Mab21 domain of cGAS comprises two lobes, separated by a deep cleft (Fig. 1b). Lobe 1 possesses the NTase fold with a two-leaved highly twisted β-sheet (β1–β8) that is flanked on the outside by two long α-helices (αA and αB). At the inner side, lining the cleft, β1 and β6 harbour the signature catalytic site residues (E200, D202, D296) of the NTase superfamily that coordinate the catalytic Mg^{2+} ions and nucleotides. Lobe 2 is a bundle of four α-helices (αE–αH), connected

[1] Department of Biochemistry and Gene Center, Ludwig-Maximilians-University, 81377 Munich, Germany. [2] Institute for Clinical Chemistry & Clinical Pharmacology, Unit for Clinical Biochemistry, University Hospital, University of Bonn, 53127 Bonn, Germany. [3] Center for Integrated Protein Sciences, 81377 Munich, Germany.
*These authors contributed equally to this work.

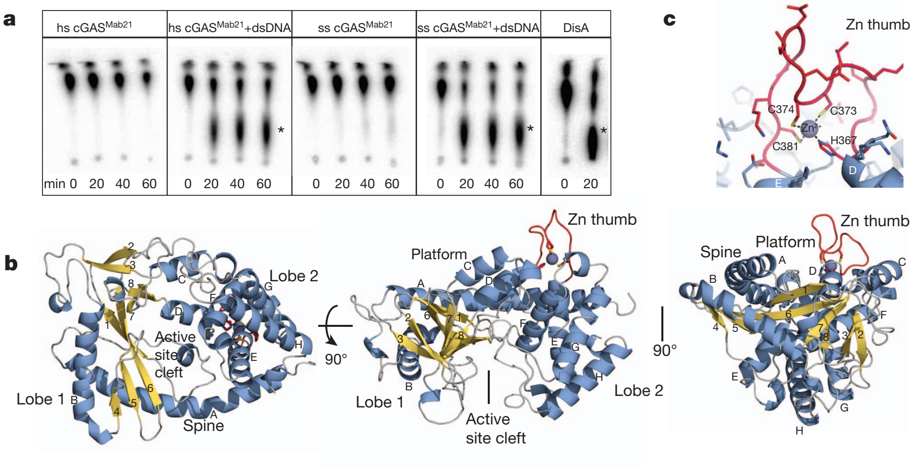


Figure 1 | Crystal Structure of cGAS^{Mab21}. **a**, Activity assays of human and porcine cGAS^{Mab21} alone or in presence of dsDNA. *Bacillus subtilis* DisA, a c-di-AMP synthase is used as positive control. The dinucleotide products are indicated with asterisks. **b**, Side and top views of cGAS^{Mab21}. The model is shown as ribbon representation with annotated domains and secondary structure (blue α-helices, yellow β-strands). **c**, Close-up view of the 'zinc thumb'.

to lobe 1 by a long 'spine' (αA), two linker helices (αC, αD) and by a long active site loop connecting αA and β1.

The molecular surface opposite the active site is a fairly flat, slightly concave 'platform', formed predominantly by αA, αC, αD and the nucleotide-binding loop. An intriguing protrusion (residues 367–382) is situated at one end of the platform. This protrusion contains highly conserved histidine and cysteines (H367, C373, C374 and C381), which together coordinate a Zn^{2+} ion (Fig. 1c). We denote this loop 'Zn thumb'. Its sequence is inserted between lobes 1 and 2 and is a highly conserved and characteristic feature of cGAS orthologues (Supplementary Fig. 1d), indicating an important functional role.

The cGAS–DNA–GTP–ATP complex

To reveal the structure of the activated conformation of cGAS, we co-crystallized cGAS$^{Mab21(td)}$ with a self-complementary 14-mer oligonucleotide, ATP, GTP and $MgCl_2$. To trap an activated conformation of cGAS^{Mab21} with DNA and bound nucleotides we mutated the NTase catalytic residues E200 and D202 to Q and N, respectively, thereby preventing catalysis during crystallization. The resulting transferase-deficient (td) variant is denoted cGAS$^{Mab21(td)}$. The structure of the cGAS$^{Mab21(td)}$–DNA–GTP–ATP complex was determined by molecular replacement using the coordinates for apo cGAS^{Mab21} as search model. $2F_o - F_c$ and $F_o - F_c$ maps revealed interpretable density for 13 out of 14 base pairs of the dsDNA duplex and for both nucleotides bound at the active site (Supplementary Fig. 2). The structure was refined at 3.1 Å resolution, resulting in a model with good *R*-factors and stereochemistry (Supplementary Table 1).

DNA is bound along the platform between the spine on one side and the Zn thumb on the other side (Fig. 2a). cGAS binds DNA predominantly by sequence-independent interactions to both phosphate-sugar backbone strands along the minor groove (Fig. 2b, c). Hereby, cGAS binds seven nucleotides at the core of the platform, which are recognized by at least eleven residues via specific side- and/or main-chain contacts. In addition to the phosphate and sugar contacts, two arginine fingers (R150 and R192) are inserted into the minor groove, additionally stabilizing the interaction in a fairly sequence-independent manner. Besides binding to the array of conserved positively charged residues at the bottom of the platform, DNA is also bound by the spine and the Zn thumb. The continuous helix of the spine in apo-cGAS^{Mab21} is interrupted in the DNA complex and a DNA backbone phosphate is bound at the central kink of the spine helix. On the other side of the platform, the Zn thumb contacts the DNA backbone near the major groove. We do not see close, direct polar contacts between Zn thumb and DNA, but do not want to rule out water-mediated interactions here (Supplementary Fig. 2a).

The Zn thumb does not substantially change conformation or location between apo and DNA-bound cGAS. It seems to be a rather rigid element, in which the zinc ion serves as a structural stabilizer of the protruding loop, similar to Zn^{2+} in regulatory domains of RIG-I-like receptors[25]. The location of the Zn thumb at the backbone near the major groove suggests that it may assist in binding to B-form DNA. In support of this, we do not see a substantial perturbation of the bound DNA from canonical B-form DNA.

Altogether, our structure suggests a specific recognition of B-form dsDNA by cGAS through an extended B-DNA binding platform and

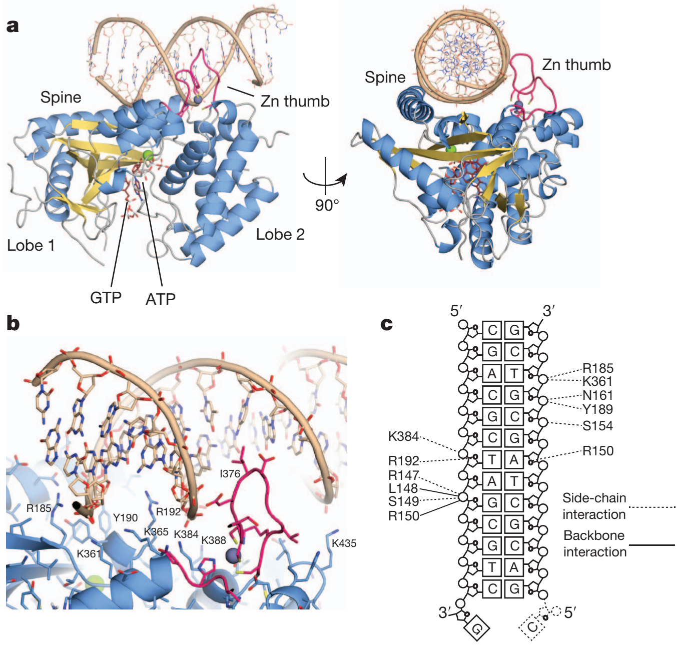


Figure 2 | The cGAS^{Mab21}–DNA–GTP–ATP complex. **a**, Side and top views of cGAS^{Mab21} (colour code of Fig. 1b) in complex with dsDNA (brown), GTP and ATP (ruby stick models). DNA binds along the platform between spine and Zn thumb. **b**, Close-up view of the DNA binding site with selected annotated residues. DNA is bound mainly via the minor groove. A notable exception is the Zn thumb near the major groove. **c**, Schematic representation of DNA–cGAS contacts.

flanking 'Zn thumb' across both lobes of the enzyme. The observed mode of binding is consistent with the key role of cGAS in sensing very different types of DNA in a sequence-independent manner[18,23].

Structure–function analysis

To validate the structural results, we mutated several conserved positively charged residues at the DNA-binding platform of human cGAS, two active site residues, two zinc ligands in the Zn thumb, or the entire Zn thumb and tested for nucleotidyl-transferase activity *in vitro* by thin-layer chromatography (TLC) (Fig. 3a). cGAS produces a product that migrates approximately in the range of c-di-AMP synthesized by DisA[26], consistent with formation of a dinucleotide. The conserved active site residues of NTases (human E225+D227; porcine E200+D202 and human G212+S213) are essential for *in vitro* activity of cGAS$^{\mathrm{Mab21}}$. Moreover, mutation of conserved positively charged residues at the centre and flanking regions of the platform (K173+R176 and K407+K411) either diminish or abolish activity, in accordance with this site being important for DNA sensing. Finally, disruption of the zinc-binding site of the thumb (human C396+C397, Zn thumbless) abolishes DNA-induced NTase activity *in vitro*, highlighting the functional importance of the conserved Zn thumb in DNA binding.

To test the effect of active site, platform and thumb mutations in living cells, we measured the transactivation of an IFN-β promoter reporter by transiently expressing human cGAS variants in HEK293T cells that stably expressed murine STING (Fig. 3b). Induction of IFN-β by cGAS$^{\mathrm{Mab21}}$ (human cGAS$^{155–552}$) in these cells is only moderately reduced compared to wild-type cGAS, showing that the Mab21 domain structurally addressed in this study is the catalytic active functional core of the sensor. The activity of full-length cGAS was abolished when residues of the NTase active site were mutated (E225Q/A+D227N/A or G212A+S213A). Mutating charged platform residues (K173A+R176A; K407A+K411A) substantially reduced the activity of cGAS in living cells. Likewise, disrupting the zinc-binding site of the thumb (C396A+C397A, Zn thumbless) severely compromised cGAS activity. These data validate the *in vitro* biochemical data and emphasize the importance of the structure-derived motifs and elements in living cells.

To see whether Zn thumb and conserved platform surface residues are important for dsDNA binding and activity, we performed electrophoretic mobility shift assays (Fig. 3c). Both porcine and human wild-type cGAS$^{\mathrm{Mab21}}$ bind efficiently to dsDNA and, surprisingly, also to dsRNA (Supplementary Fig. 3a, c). The mutations in platform and thumb either did not affect DNA/RNA binding under these conditions, or reduced but did not abolish it (Supplementary Fig. 3b). However, both mutants fail to show DNA-stimulated activity under conditions where they still bind DNA, and dsRNA fails to stimulate activity under conditions where it binds robustly to the protein (Supplementary Fig. 3c, d). Thus, although these analyses validate the functional relevance of the DNA binding platform and Zn thumb on activating cGAS, they suggest that DNA or RNA interactions per se are not sufficient to activate the enzyme, indicating for instance the necessity for a precise DNA-induced structural switch.

NTase and DNA induced structural switch

To reveal the mechanism of activation of cGAS by DNA, we first analysed the NTase mechanism. We see clear electron density for two nucleotide triphosphate moieties (Supplementary Fig. 2b). The two bases partially stack in an approximately 90° rotated orientation and inserted into a hydrophobic/aromatic pocket, sandwiched between

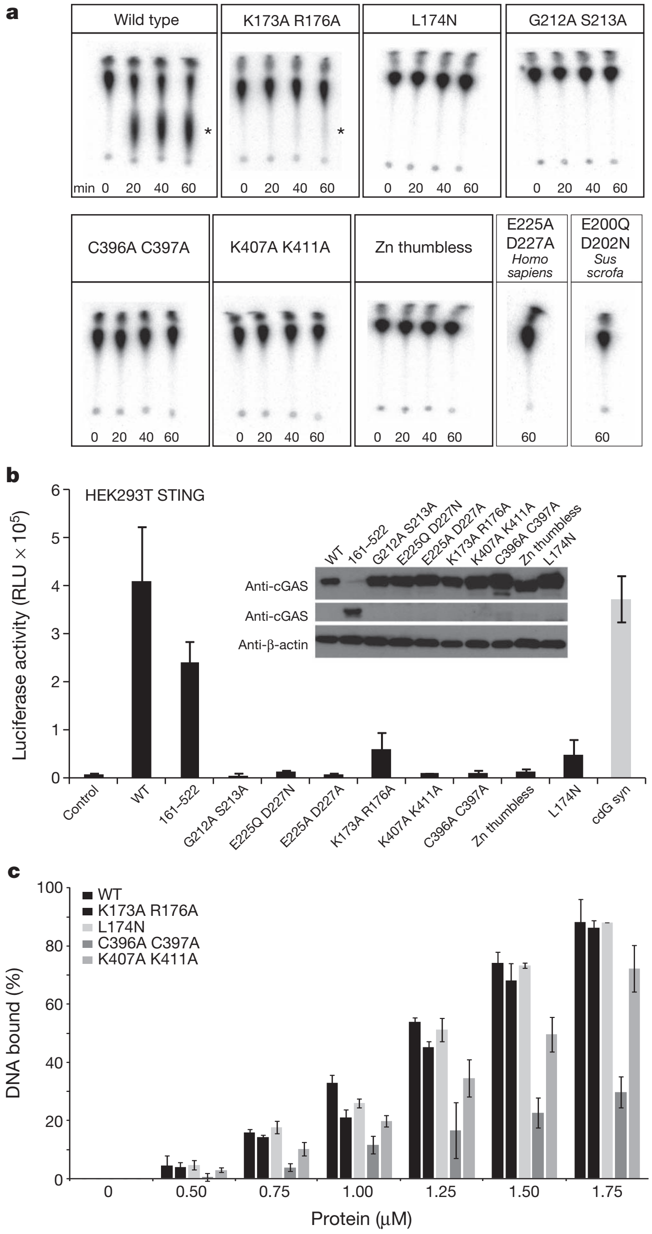


Figure 3 | Platform and Zn thumb are involved in dsDNA-dependent activity. a, NTase assays performed with different cGAS$^{\mathrm{Mab21}}$ mutants (2 μM) in presence of 3 μM dsDNA (50-mer). Human wild-type cGAS$^{\mathrm{Mab21}}$ (positive control) synthesizes dinucleotide, DNA binding site mutant K173A+R176A show reduced activity. K407A+K411A DNA binding site mutant, C396A+C397A Zn thumb mutant, Zn thumbless, L174N structural switch mutant, active site mutants E200Q+D202N of porcine cGAS$^{\mathrm{Mab21}}$ and E225A+D227A and G212A+S213A of human cGAS$^{\mathrm{Mab21}}$ are inactive. The asterisk indicates the dinucleotide product. **b**, IFN-β stimulation of cGAS mutants in HEK293T cells stably expressing murine STING. HEK293T cells were transfected with plasmids encoding indicated constructs along with the IFN-β promoter reporter plasmid pIFN-β-GLUC. Luciferase activity is plotted: mean ± s.d. ($n = 3$). Both full-length and the crystallized region (cGAS$^{\mathrm{Mab21}}$ human 155–522) induce IFN-β promoter transactivation. Active site mutations (G212A+S213A and E225Q/A+D227N/A) abolish IFN-β stimulation. DNA-binding site mutants (K173A+R176A, K407A+K411A), Zn thumb mutants (C396A+C397A, Zn thumbless) and structural switch mutant (L174N) either reduce or abolish IFN-β stimulation. Empty vector was used as negative control whereas cyclic-di-GMP synthase (cdG syn) expressing vector was used as positive control. Inset: western blot showing wild-type and mutant protein levels with β-actin as loading control. **c**, Electrophoretic mobility shift analysis of 50-mer dsDNA (0.2 μM) bound to cGAS$^{\mathrm{Mab21}}$ mutants at indicated concentrations. Plotted bars, mean ± s.d. ($n = 3$). Whereas K407A+K411A DNA binding site mutant and C396A+C397A Zn thumb mutant show slightly reduced but not impaired affinity to dsDNA, no detectable binding change was observed with the other mutants.

I298 (lobe 1) and Y413 (lobe 2). The current resolution of the diffraction data does not allow us to unambiguously determine which base is adenine and which guanine. Binding of R353 at nucleobase 1 (the 'receiving substrate' of NTases) near O6 and N7 would argue for this being guanine. In general, nucleobase 1 (interpreted as guanine here) is in hydrogen bonding distance to S355, S357 and T186, suggesting that this nucleotide is specifically recognized. In contrast, we do not observe direct hydrogen-bonding contacts of the protein to nucleobase 2 (the 'transferred' nucleotide in NTases; interpreted as adenine here). Nevertheless, this recognition might be mediated via water molecules such as in 3′ terminal uridylyl transferases[27].

The structure provides a mechanism for attack of nucleotide 1 on nucleotide 2, consistent with the mechanism of other NTases, for example, CCA adding enzyme[28]. The triphosphate chain of nucleotide 2 is well coordinated via S188 (lobe 1), S412 (lobe 2) and Mg^{2+} bound to E200 (Q in cGAS$^{\text{Mab21(td)}}$) and D202 (N in cGAS $^{\text{Mab21(td)}}$). As a consequence, the relative orientation of lobes 1 and 2 is important for the phosphate coordination of nucleotide 2. In our conformation, the α-phosphate of nucleotide 2 is well placed and oriented to promote nucleophilic attack of the sugar 2′ OH from nucleotide 1 to form the 2′-5′ linkage (Fig. 4a, see ref. 29). The attacking OH of nucleotide 1 is polarized and activated by D296, consistent with the conserved features of NTases[24]. A second Mg^{2+} could be important for this catalytic step. However, distinct localization will require higher resolution.

cGAS is proposed to form a cyclic-dinucleotide, which would require a second catalysis step and an additional attack of the OH of nucleotide 2 on the phosphate of nucleotide 1. Such an attack will require an almost 180° flip of the sugar moiety of nucleotide 2 to place its α-phosphate appropriately. In principle this is possible: in the course of our studies we determined the crystal structure of cGAS$^{\text{Mab21}}$ bound to UTP in the absence of DNA and do observe an appropriate flip of the sugar moiety (Supplementary Fig. 4). In any case, our structure satisfactorily explains the catalysis of formation of a specific (at present linear) dinucleotide by cGAS, but formation of a cyclic dinucleotide needs to be addressed in future studies.

To reveal a potential activation mechanism of cGAS, we superimposed apo-cGAS, cGAS$^{\text{Mab21}}$–UTP and cGAS$^{\text{Mab21(td)}}$–DNA–GTP–ATP complex (Fig. 4b, c and Supplementary Fig. 5a, b). We used cGAS$^{\text{Mab21}}$–UTP because UTP binding orders the β-sheets on lobe 1 and we can also visualize conformational changes specifically induced by dsDNA rather than the nucleotides.

Although UTP binding to cGAS ordered to some extend the nucleotide-binding loop in the active site, it did not substantially change the overall structure and active site geometry of cGAS (Supplementary Fig. 5b). In contrast, DNA phosphate binding to the spine (Fig. 4b) triggers a substantial structural switch in the spine helix (Fig. 4c) that closes lobes 1 and 2 and rearranges the active site loop, allowing magnesium coordinating of E200 to position and activate nucleotide 2.

To test the role of this DNA-induced structural switch we mutated human L174 to N. L174 (porcine L148) is repositioned in response to DNA binding to stabilize the nucleotide-binding loop, but does not directly bind DNA or NTPs (Supplementary Fig. 5c). Although L174N shows fairly normal DNA binding (Fig. 3c and Supplementary Fig. 3b), it lacks DNA-stimulated cGAMP synthetase activity *in vitro* (Fig. 3a) and shows decreased interferon stimulation in cells (Fig. 3b). Thus, the structural and biochemical data suggest that cGAS is activated by a DNA-induced structural switch that rearranges the NTase active site.

Conclusion

Here we provide the structure and mechanism of activation of the cytosolic DNA sensor cyclic-GMP-AMP synthase that readily explain the synthesis of a linear dinucleotide intermediate by cGAS in response to DNA binding. The backbone binding of a canonical B-DNA by cGAS is consistent with a broad specificity innate immune PRR for cytosolic DNA and the structural elements of cGAS such as the positioning of residues involved in minor-groove binding, arginine fingers and the Zn thumb suggest that cGAS specifically responds to B-form DNA. This might explain the function of other innate immune DNA sensors to detect non-canonical DNA structures, such as DAI[5]. A structural switch transmitted by proper B-form DNA binding to the active site could also explain the lack of activation by dsRNA or in mutants that still bind DNA: slightly different conformations of RNA-bound or DNA-bound mutant cGAS would not trigger robust cGAMP synthesis as even small differences in the active site geometry can strongly affect catalytic rates of enzymes.

In future, it will be important to address the specificity for other DNA structures in the activation of cGAS in more detail to see which types of DNA structures can activate cGAS. It will also be important to investigate additional requirements for efficient DNA sensing *in vivo*, because although shorter dsDNA molecules can stimulate cGAS$^{\text{Mab21}}$ *in vitro*, DNA larger than 50-mer is required for efficient IFN stimulation *in vivo*[14,19]. One possibility is that fraying of shorter DNA molecules prevents efficient stimulation or that the positively charged N terminus contributes to sensing of longer DNA molecules. In addition, STING might have a direct role in DNA binding in a larger context *in vivo*[19], although we do not see strong DNA binding *in vitro* and IFN stimulation in response to DNA in HEK293T cells in the absence of cGAS (Supplementary Fig. 6).

Interestingly, cGAS has remarkable fold similarity to the antiviral protein oligoadenylate synthase 1 (OAS1)[30,31] (Fig. 5). OAS1 synthesizes 2′-5′ linked oligoadenylate chains in response to binding to

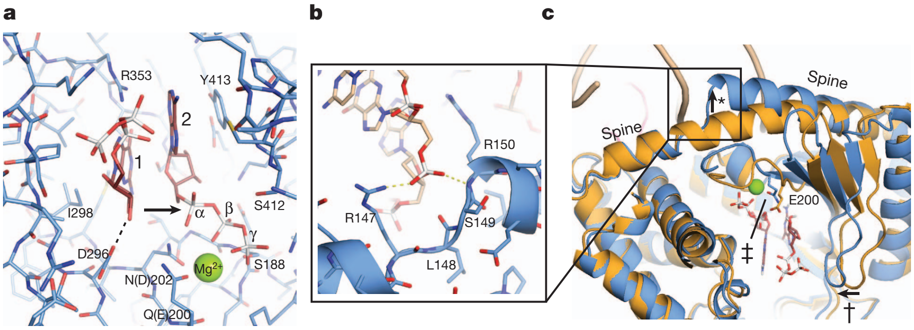


Figure 4 | NTase and DNA-induced structural switch. a, Close-up view of the NTase active site. Selected residues that are implicated in binding and catalysis are annotated. Both base moieties partially stack to each other and are further bound by stacking to Y413 and recognition by R353. E200 (mutated to Q in our structure) and D202 (mutated to N in our structure) bind an active site magnesium that coordinates phosphates of nucleotide 2. The attacking OH of nucleotide 1 is activated by D296 for nucleophilic attack on the α-phosphate of nucleotide 2 (arrow). **b**, Close-up view of DNA backbone phosphate binding at the spine. **c**, This DNA phosphate binding triggers a change in the spine helix (*), which allows a closure of the active site cleft (†) and repositioning of the substrate binding loop for Mg^{2+} coordination of E200 (‡).

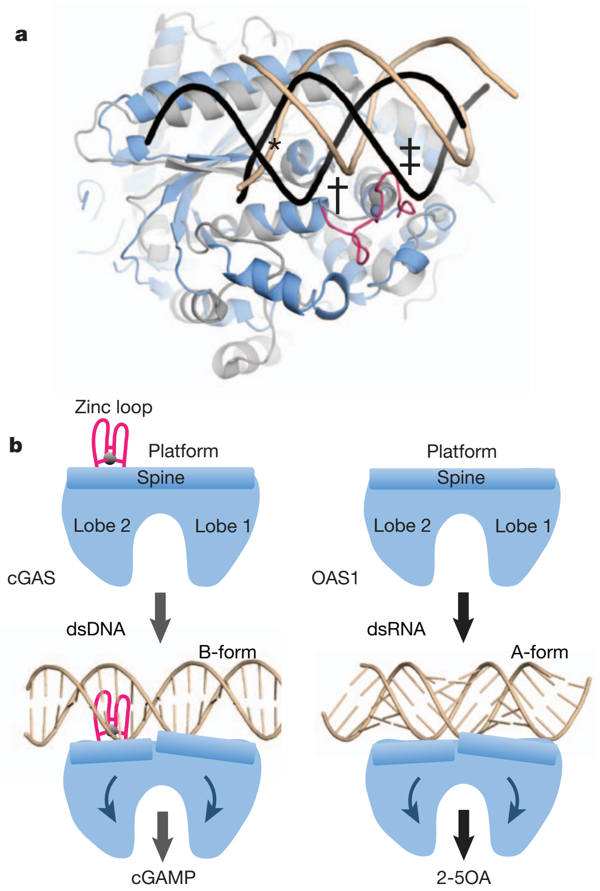


Figure 5 | Model for DNA sensing by cGAS. a, Superposition of cGAS–DNA (blue) with OAS1–RNA (grey) shows key elements for nucleic duplex selectivity. Both enzymes bind one DNA (brown)/RNA (black) backbone at the same protein site (*). The Zn thumb specifically recognizes the position of the second DNA strand in B-form (†).However, it would clash with A-form RNA/DNA (‡). **b**, Unified activation model for cytosolic double-stranded nucleic acid sensing by cGAS and OAS1 NTases by a ligand induced structural switch. 2-5OA, 2′-5′ linked oligoadenylate chains.

cytosolic dsRNA. The structural similarity not only embraces the overall fold, several active site features and arrangement of lobes 1 and 2, but also certain structural elements of the platform, including the long 'spine' helix. Like cGAS, OAS1 binds dsRNA along the 'platform' and triggers a structural change that is transmitted to the active site[31]. However, whereas OAS1 is activated by A-form RNA, cGAS is activated by B-form DNA. The Zn thumb in cGAS, missing in OAS1, probably acts as a molecular 'ruler' to specifically trigger activation in response to B-form but not A-form nucleic acids (Fig. 5). Despite these differences, cGAS shows a structural switch induced by dsDNA that is very similar to that of OAS1 induced by dsRNA[31] (Fig. 5). Thus, our results structurally unify dsDNA and dsRNA sensing by cGAS and OAS1 NTases, respectively, in the innate immune system and suggest that both processes are evolutionarily connected.
Note added in proof: After submission of the revised version of this manuscript, Gao *et al.*[32] reported related structures of cGAS and its complexes with DNA and nucleotides.

METHODS SUMMARY

Proteins were produced in *Escherichia coli* and purified by affinity, ion exchange and size exclusion chromatography. Apo, UTP- and DNA–ATP–GTP-bound cGAS$^{\text{Mab21}}$ and its catalytic inactive form were crystallized by hanging or sitting drop vapour diffusion. The structure of apo cGAS$^{\text{Mab21}}$ was determined by single-anomalous dispersion phasing on selenomethionine derivatized protein. The other structures were determined by molecular replacement using apo cGAS$^{\text{Mab21}}$ as search model. NTase assays were performed by thin layer chromatography and phosphor imaging. DNA and RNA binding were assessed by electrophoretic mobility shift assays. Analysis of cGAS mutants in living cells were performed in HEK 293T cells stably expressing full-length murine STING and transfected with an IFN-β promoter reporter plasmid.

Full Methods and any associated references are available in the online version of the paper.

Received 21 March; accepted 20 May 2013.
Published online 30 May 2013.

1. Rathinam, V. A. K. & Fitzgerald, K. A. Cytosolic surveillance and antiviral immunity. *Curr. Opin. Virol.* **1,** 455–462 (2011).
2. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140,** 805–820 (2010).
3. Keating, S. E., Baran, M. & Bowie, A. G. Cytosolic DNA sensors regulating type I interferon induction. *Trends Immunol.* **32,** 574–581 (2011).
4. Krug, A. Nucleic acid recognition receptors in autoimmunity. *Handb. Exp. Pharmacol.* **183,** 129–151 (2008).
5. Takaoka, A. *et al.* DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response. *Nature* **448,** 501–505 (2007).
6. Bürckstümmer, T. *et al.* An orthogonal proteomic-genomic screen identifies AIM2 as a cytoplasmic DNA sensor for the inflammasome. *Nature Immunol.* **10,** 266–272 (2009).
7. Fernandes-Alnemri, T., Yu, J. W., Datta, P., Wu, J. & Alnemri, E. S. AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. *Nature* **458,** 509–513 (2009).
8. Hornung, V. *et al.* AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature* **458,** 514–518 (2009).
9. Ablasser, A. *et al.* RIG-I-dependent sensing of poly(dA:dT) through the induction of an RNA polymerase III-transcribed RNA intermediate. *Nature Immunol.* **10,** 1065–1072 (2009).
10. Chiu, Y. H., Macmillan, J. B. & Chen, Z. J. RNA polymerase III detects cytosolic DNA and induces type I interferons through the RIG-I pathway. *Cell* **138,** 576–591 (2009).
11. Yang, P. *et al.* The cytosolic nucleic acid sensor LRRFIP1 mediates the production of type I interferon via a β-catenin-dependent pathway. *Nature Immunol.* **11,** 487–494 (2010).
12. Kim, T. *et al.* Aspartate-glutamate-alanine-histidine box motif (DEAH)/RNA helicase A helicases sense microbial DNA in human plasmacytoid dendritic cells. *Proc. Natl Acad. Sci. USA* **107,** 15181–15186 (2010).
13. Zhang, Z. *et al.* The helicase DDX41 senses intracellular DNA mediated by the adaptor STING in dendritic cells. *Nature Immunol.* **12,** 959–965 (2011).
14. Unterholzner, L. *et al.* IFI16 is an innate immune sensor for intracellular DNA. *Nature Immunol.* **11,** 997–1004 (2010).
15. Rathinam, V. A. & Fitzgerald, K. A. Innate immune sensing of DNA viruses. *Virology* **411,** 153–162 (2011).
16. Ishikawa, H., Ma, Z. & Barber, G. N. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* **461,** 788–792 (2009).
17. Ishikawa, H. & Barber, G. N. STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. *Nature* **455,** 674–678 (2008).
18. Wu, J. *et al.* Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA. *Science* **339,** 826–830 (2013).
19. Abe, T. *et al.* STING recognition of cytoplasmic DNA instigates cellular defense. *Mol. Cell* **50,** 5–15 (2013).
20. Burdette, D. L. *et al.* STING is a direct innate immune sensor of cyclic di-GMP. *Nature* **478,** 515–518 (2011).
21. McWhirter, S. M. *et al.* A host type I interferon response is induced by cytosolic sensing of the bacterial second messenger cyclic-di-GMP. *J. Exp. Med.* **206,** 1899–1911 (2009).
22. Woodward, J. J., Iavarone, A. T. & Portnoy, D. A. c-di-AMP secreted by intracellular *Listeria monocytogenes* activates a host type I interferon response. *Science* **328,** 1703–1705 (2010).
23. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* **339,** 786–791 (2013).
24. Kuchta, K., Knizewski, L., Wyrwicz, L. S., Rychlewski, L. & Ginalski, K. Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. *Nucleic Acids Res.* **37,** 7701–7714 (2009).
25. Cui, S. *et al.* The C-terminal regulatory domain is the RNA 5′-triphosphate sensor of RIG-I. *Mol. Cell* **29,** 169–179 (2008).
26. Witte, G., Hartung, S., Buttner, K. & Hopfner, K. P. Structural biochemistry of a bacterial checkpoint protein reveals diadenylate cyclase activity regulated by DNA recombination intermediates. *Mol. Cell* **30,** 167–178 (2008).
27. Stagno, J., Aphasizheva, I., Rosengarth, A., Luecke, H. & Aphasizhev, R. UTP-bound and apo structures of a minimal RNA uridylyltransferase. *J. Mol. Biol.* **366,** 882–899 (2007).
28. Xiong, Y. & Steitz, T. A. Mechanism of transfer RNA maturation by CCA-adding enzyme without using an oligonucleotide template. *Nature* **430,** 640–645 (2004).
29. Ablasser, A. *et al.* cGAS produces a 2′-5′-linked cyclic dinucleotide second messenger that activates STING *Nature* http://dx.doi.org/10.1038/nature12306 (30 May 2013).

30. Hartmann, R., Justesen, J., Sarkar, S. N., Sen, G. C. & Yee, V. C. Crystal structure of the 2′-specific and double-stranded RNA-activated interferon-induced antiviral protein 2′-5′-oligoadenylate synthetase. *Mol. Cell* **12,** 1173–1185 (2003).
31. Donovan, J., Dufner, M. & Korennykh, A. Structural basis for cytosolic double-stranded RNA surveillance by human oligoadenylate synthetase 1. *Proc. Natl Acad. Sci. USA* **110,** 1652–1657 (2013).
32. Gao, P. *et al.* Cyclic [G(2′,5′)pA(3′,5′)p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. *Cell* **153,** 1094–1107 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Butryn for comments on the manuscript. We thank the Max-Planck-Crystallization facility for initial crystal screening and the Swiss Light Source, European Synchrotron Radiation Facility and the German electron synchrotron Petra III for beam time and on-site assistance. This work was funded by the National Institutes of Health (U19AI083025), the European Research Council Advanced Grant 322869, and the Center for Integrated Protein Science Munich (CIPSM) to K.-P.H., by DFG grant 3717/2-1 to G.W., by GRK1721 to K.-P.H. and G.W., by DFG grant SFB670 and ERC grant 243046 to V.H.; C.C.O.M. is supported by GRK1721.

Author Contributions F.C. crystallized and determined the structure of cGAS, performed biochemical assays, interpreted data and wrote the manuscript. T.D. crystallized and refined the DNA complex. C.C.O.M., A.A., T.D. and M.M. performed biochemical assays. G.W. performed biochemical assays, interpreted data and helped with structure determination. V.H. supervised the cell-based experiments and interpreted data. K.-P.H designed the research, helped with structure determination, interpreted data and wrote the manuscript.

Author Information Coordinates and structure factors have been deposited at the Protein Data Bank (4JLX, 4JLZ and 4KB6). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.-P.H. (hopfner@genzentrum.lmu.de).

METHODS

Constructs and cloning. The sequence encoding full-length or truncated *Homo sapiens* and *Sus scrofa* cGAS were amplified from total cDNA (courtesy of S. Bauersachs) and cloned into pIRESneo3 (Clontech) or a modified pET21 (Novagen), respectively. The mutants were generated by site-directed mutagenesis using PfuUltra (Stratagene). Zn thumbless mutant was created by replacing residues 390–405 (*Homo sapiens*) by three Gly-Ser replicates.

Protein production and purification. All proteins were produced in *E. coli* Rosetta (DE3) or B834 (DE3) strains for native or selenomethionine derivative proteins, respectively. Bacteria were grown until a D_{600} of 0.6 to 0.8 was reached and expression was induced at 18 °C for 16 to 18 h with 0.1 mM IPTG. Proteins were purified by Ni-NTA agarose resin and incubated with tobacco etch virus (TEV) protease (ratio 1:50) at 4 °C overnight to remove the 6xHis-MBP-tag. The proteins were further purified by cation exchange chromatography followed by size exclusion chromatography using a Superdex 200 column (GE Healthcare), equilibrated in 20 mM Tris pH 7.5, 150 mM NaCl and 1 mM DTT. Purified proteins were concentrated to 10 mg ml^{-1} for crystallization. Human STING 139–379 was purified as described[33]. All purified proteins were frozen in liquid N_2 and stored at −80 °C.

Crystallization of $cGAS^{Mab21}$. Purified porcine cGAS (10 mg ml^{-1}) was crystallized by hanging drop vapour diffusion in 20% PEG3350 and 200 mM sodium malonate. The crystals appeared after one day at 20 °C and were flash frozen after addition of glycerol to a final concentration of 15% (v/v). The selenomethionine derivatized protein was crystallized in 100 mM Bis-Tris propane pH 6.3, 18% PEG3350 and 200 mM sodium malonate and cryo protected with 20% ethane-1,2-diol before flash freezing. UTP bound crystals were obtained by adding 20 mM $MgCl_2$ and 1:10 (v/v) of 50 mM of nucleotide in 100 mM Tris pH 7.5 to the protein before crystallization.

For crystallizing the DNA–GTP–ATP–cGAS complex 20 mM $MgCl_2$, 2 mM of both nucleotides and 14 bp dsDNA (5′-CGACGCTAGCGTCG-3′) in a molar ratio of 1:1.2 protein:DNA were added to the inactive porcine $cGAS^{Mab21(td)}$ (E200Q+D202N) (10 mg ml^{-1}). Crystals were obtained by hanging drop vapour diffusion in 50 mM sodium cacodylate pH 7.0, 2.5 mM spermine, 60 mM $MgCl_2$ and 3% (v/v) PEG 400 after one day at 20 °C. The crystals were soaked in reservoir solution containing 25% (v/v) glycerol before flash freezing.

Data collection and refinement. X-ray diffraction data of cGAS and cGAS-UTP were collected at X06SA beamline (Swiss Light Source, Switzerland) and diffraction data of the $cGAS^{Mab21(td)}$–GTP–ATP–DNA complex were collected at PetraIII beamline P14 (EMBL/DESY, Hamburg, Germany) at 100 K. The selenomethionine derivative data were collected at the selenium peak wavelength ($\lambda = 0.97961$ Å). Data processing was carried out with XDS[34]. AutoSHARP was used to locate Se sites (SAD data set) and to produce an initial solvent flattened map[35]. An initial model was built using iterative cycles of Buccaneer[36] and ARP/wARP classic[37]. The model was optimized by alternating manual building with Coot[38] and refinement using Phenix[39] against a 2.0 Å native data set. The structure of UTP-bound cGAS and the DNA–GTP–ATP–cGAS complex structure were determined using molecular replacement with Phaser[40] and optimized by manual building with Coot and refinement with Phenix or Autobuster[41]. Data collection and refinement statistics are listed in Supplementary Table 1.

NTase assays. NTase assays were performed as described in ref. 26. Reaction mixtures with the indicated concentrations of protein and DNA (40-mer: 5′-GGATACGTAACAACGCTTATGCATCGCCGCCGCTACATCC-3′, 50-mer: 5′-GGATACGTAACAACGCTTATGCATCGCCGCCGCTACATCCCTGAGCTGAC-3′) (unless indicated 50-mer dsDNA is used) or RNA (sequence as 50-mer DNA) in 0.1 M NaCl, 40 mM Tris pH 7.5 and 10 mM $MgCl_2$ were started by addition of 100 μM ATP and 100 μM GTP containing 1:600 [α^{32}P]ATP and/or [α^{32}P]GTP (3,000 Ci $mmol^{-1}$, Hartmann Analytic). Analysis of the reaction products was done using thin layer chromatography (PEI-Cellulose F plates, Merck) with 1 M $(NH_4)_2SO_4$/1.5 M KH_2PO_4 pH 3.8 as running buffer for the TLC plates. Assays were performed at 35 °C. The dried TLC plates were analysed by phosphor imaging (GE Healthcare).

Electrophoretic mobility shift assays. 0.2 μM of dsDNA or dsRNA (same sequences used for NTase assays) was incubated with indicated amount of purified protein for 30 min on ice. As reaction buffer 20 mM Tris pH 8.0 and 200 mM NaCl was used. Samples were separated by 1% agarose gel prepared with Gel-Red (Biotium) as suggested by the manufacturer. The gel images were analysed using ImageJ.

Reporter assays. HEK 293T cells stably expressing full-length murine STING (2×10^4 cells in each well of a 96-well plate) were transiently transfected with 25 ng IFN-β promoter reporter plasmid (pIFN-β-GLUC) in conjunction with 200 ng cGAS expression vectors using GeneJuice (Novagen) as indicated by the manufacturer. A codon-optimized version the diguanylate cyclase domain (83–248) of TM1788 (*Thermotoga maritima* MSB8) harbouring a point mutation (R158A) to enhance c-di-GMP production was cloned into pEFBOS to contain a carboxy-terminal haemagglutinin (HA) tag[42]. This construct (c-di-GMP-synthase) was used to induce c-di-GMP production within 293T cells upon transient overexpression, which served as positive control. 14 h post transfection luciferase activity was assessed.

THP-1 cells were stimulated with 200 ng of either 50-mer dsDNA (as in NTase assays) or tri-phosphate-RNA complexed with Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Supernatants were collected 18 h after stimulation and assayed for IP-10 production via ELISA. 90-mer DNA used is as described in ref. 19. CMA was purchased from Sigma Aldrich.

Immunoblotting. Cells were lysed in 1× Laemmli buffer and denatured at 95 °C for 5 min. Probes were separated by 10% SDS–PAGE and transferred onto nitrocellulose membranes. Blots were incubated with anti-cGAS (Sigma Aldrich), anti-phospho-IRF3 (Cell Signaling Technology) or anti-β-actin-IgG–horseradish peroxidase (HRP). Goat anti-rabbit-IgG–HRP was purchased from Santa Cruz Biotechnology.

33. Cavlar, T., Deimling, T., Ablasser, A., Hopfner, K. P. & Hornung, V. Species-specific detection of the antiviral small-molecule compound CMA by STING. *EMBO J.* **32,** 1440–1450 (2013).
34. Kabsch, W. XDS. *Acta Crystallogr.* **66,** 125–132 (2010).
35. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. Automated structure solution with autoSHARP. *Methods Mol. Biol.* **364,** 215–230 (2007).
36. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr.* **62,** 1002–1011 (2006).
37. Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP's model-building algorithms. I. The main chain. *Acta Crystallogr.* **58,** 968–975 (2002).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr.* **60,** 2126–2132 (2004).
39. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* **66,** 213–221 (2010).
40. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
41. Blanc, E. *et al.* Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr.* **60,** 2210–2221 (2004).
42. Rao, F. *et al.* Enzymatic synthesis of c-di-GMP using a thermophilic diguanylate cyclase. *Anal. Biochem.* **389,** 138–142 (2009).

doi:10.1038/nature12184

The rapid assembly of an elliptical galaxy of 400 billion solar masses at a redshift of 2.3

Hai Fu[1], Asantha Cooray[1], C. Feruglio[2], R. J. Ivison[3], D. A. Riechers[4], M. Gurwell[5], R. S. Bussmann[5], A. I. Harris[6], B. Altieri[7], H. Aussel[8], A. J. Baker[9], J. Bock[10,11], M. Boylan-Kolchin[1], C. Bridge[11], J. A. Calanog[1], C. M. Casey[12], A. Cava[13], S. C. Chapman[14], D. L. Clements[15], A. Conley[16], P. Cox[2], D. Farrah[17], D. Frayer[18], R. Hopwood[15], J. Jia[1], G. Magdis[19], G. Marsden[20], P. Martínez-Navajas[21,22], M. Negrello[23], R. Neri[2], S. J. Oliver[24], A. Omont[25], M. J. Page[26], I. Pérez-Fournon[21,22], B. Schulz[27], D. Scott[20], A. Smith[24], M. Vaccari[28], I. Valtchanov[7], J. D. Vieira[11], M. Viero[11], L. Wang[24], J. L. Wardlow[1] & M. Zemcov[10]

Stellar archaeology[1] shows that massive elliptical galaxies formed rapidly about ten billion years ago with star-formation rates of above several hundred solar masses per year. Their progenitors are probably the submillimetre bright galaxies[2] at redshifts *z* greater than 2. Although the mean molecular gas mass[3] (5×10^{10} solar masses) of the submillimetre bright galaxies can explain the formation of typical elliptical galaxies, it is inadequate to form elliptical galaxies[4] that already have stellar masses above 2×10^{11} solar masses at $z \approx 2$. Here we report multi-wavelength high-resolution observations of a rare merger of two massive submillimetre bright galaxies at $z = 2.3$. The system is seen to be forming stars at a rate of 2,000 solar masses per year. The star-formation efficiency is an order of magnitude greater than that of normal galaxies, so the gas reservoir will be exhausted and star formation will be quenched in only around 200 million years. At a projected separation of 19 kiloparsecs, the two massive starbursts are about to merge and form a passive elliptical galaxy with a stellar mass of about 4×10^{11} solar masses. We conclude that gas-rich major galaxy mergers with intense star formation can form the most massive elliptical galaxies by $z \approx 1.5$.

HXMM01 (1HERMES S250 J022016.5−060143) was identified as an unusually bright submillimetre bright galaxy (SMG)[5] in the Herschel Multi-tiered Extragalactic Survey (HerMES[6]). We observed it with a variety of ground-based telescopes from optical to radio to obtain higher-resolution images and to improve the sampling of the spectral energy distribution. The source is resolved into two similarly bright components at $z \approx 2.308$ separated by 3″ (Fig. 1; hereafter X01N and X01S for the northern and southern component, respectively), which are connected by a bridge of material reminiscent of tidal tails frequently seen in galaxy mergers. The carbon monoxide (CO) $J = 1 \rightarrow 0$, $J = 4 \rightarrow 3$, and Hα spectra all show slightly different redshifts for these two components. The velocity separation is $260 \pm 70\ \mathrm{km\ s^{-1}}$ and $330 \pm 220\ \mathrm{km\ s^{-1}}$ from the CO $J = 1 \rightarrow 0$ and Hα spectra, respectively. In addition, the CO $J = 1 \rightarrow 0$ linewidths for the two components are significantly different ($970 \pm 150\ \mathrm{km\ s^{-1}}$ versus $660 \pm 100\ \mathrm{km\ s^{-1}}$ in full width at half-maximum), establishing clearly that they are two distinct galaxies undergoing a merger. HXMM01 is close to two low-redshift galaxies and is weakly gravitationally magnified. From our lens model (see Supplementary Information), we determined that the magnification factors are 1.8 ± 0.5, 1.4 ± 0.2, and 1.6 ± 0.3 for X01N, X01S, and the entire system, respectively. Here we quote lensing-corrected values, and the uncertainties in magnification have been propagated into their errors.

With a flux density of 20 ± 4 mJy at 850 μm, HXMM01 is among the brightest SMGs known; SMGs typically have a flux density of 5 mJy at 850 μm. The spectral energy distribution between 160 μm and 2.1 mm is fully consistent with a modified black body with a characteristic dust temperature of 55 ± 3 K. The dust temperature is much hotter than that of normal star-forming galaxies, and lies at the higher end of that for starbursts[7]. We determine a total infrared (8–1,000 μm) luminosity of $L_{\mathrm{IR}} = (2.0 \pm 0.4) \times 10^{13} L_{\odot}$ for all components combined, where $L_{\odot}$ is the solar luminosity. This luminosity implies an instantaneous star-formation rate[8] (SFR) of 2,000 ± 400 solar masses per year ($M_{\odot}\ \mathrm{yr}^{-1}$) for a Chabrier initial mass function.

We estimate the total molecular gas mass from the CO $J = 1 \rightarrow 0$ line luminosity. The area–velocity-integrated CO $J = 1 \rightarrow 0$ brightness temperature, which is also a luminosity measure, is $(2.9 \pm 0.7) \times 10^{11}\ \mathrm{K\ km\ s^{-1}\ pc^{2}}$, implying a mass of $(2.3 \pm 0.6) \times 10^{11} M_{\odot}$ of molecular gas using our derived conversion factor (see Supplementary Information). To our knowledge, this is the highest-known intrinsic CO $J = 1 \rightarrow 0$ line luminosity for an SMG (Fig. 2). Table 1 lists the known SMG mergers that are well separated into two galaxies. It shows that HXMM01 is the brightest, most luminous, and most gas-rich SMG merger that is known.

We estimate stellar masses of $(9 \pm 2) \times 10^{10} M_{\odot}$ and $(1.3 \pm 0.3) \times 10^{11} M_{\odot}$ for X01N and X01S respectively, by modelling their ultraviolet-to-millimetre spectral energy distributions using the public code MAGPHYS[9]. HXMM01 is hence a merger of two massive galaxies in terms of both stellar and gas content. The combined gas-to-baryon fraction of HXMM01 is $(52 \pm 5)\%$, far above the mean gas-to-baryon fraction of star-forming galaxies with stellar masses above $10^{11} M_{\odot}$ at $z \approx 2$ (Fig. 3). Massive galaxies with such a high gas fraction are not reproduced in cosmological simulations[10]. HXMM01 may represent a

[1] Department of Physics and Astronomy, University of California, Irvine, California 92697, USA. [2] Institut de RadioAstronomie Millimétrique, 300 Rue de la Piscine, Domaine Universitaire, 38406 Saint Martin d'Hères, France. [3] UK Astronomy Technology Centre, Royal Observatory, Edinburgh EH9 3HJ, UK. [4] Department of Astronomy, Cornell University, 610 Space Science Building, Ithaca, New York 14853, USA. [5] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [6] Department of Astronomy, University of Maryland, College Park, Maryland 20742-2421, USA. [7] Herschel Science Centre, European Space Astronomy Centre, Villanueva de la Cañada, 28691 Madrid, Spain. [8] Laboratoire AIM-Paris-Saclay, CEA/DSM-CNRS-Université Paris Diderot, Irfu/SAp, CEA-Saclay, 91191 Gif-sur-Yvette Cedex, France. [9] Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, New Jersey 08854, USA. [10] Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA. [11] California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. [12] Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA. [13] Departamento de Astrofísica, Facultad de Ciencias Físicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain. [14] Department of Physics and Atmospheric Science, Dalhousie University, 6310 Coburg Road, Halifax B3H 4R2, Canada. [15] Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK. [16] Center for Astrophysics and Space Astronomy 389-UCB, University of Colorado, Boulder, Colorado 80309, USA. [17] Department of Physics, Virginia Tech, Blacksburg, Virginia 24061, USA. [18] National Radio Astronomy Observatory (NRAO), PO Box 2, Green Bank, West Virginia 24944, USA. [19] Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK. [20] Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, British Columbia V6T 1Z1, Canada. [21] Instituto de Astrofísica de Canarias (IAC), E-38200 La Laguna, Tenerife, Spain. [22] Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38205 La Laguna, Tenerife, Spain. [23] INAF—Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122 Padova, Italy. [24] Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK. [25] Institut d'Astrophysique de Paris, UMR 7095, CNRS, UPMC Université Paris 06, 98bis boulevard Arago, F-75014 Paris, France. [26] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK. [27] Infrared Processing and Analysis Center, MS 100-22, California Institute of Technology, Jet Propulsion Laboratory, Pasadena, California 91125, USA. [28] Astrophysics Group, Physics Department, University of the Western Cape, Private Bag X17, Bellville 7535, Cape Town, South Africa.

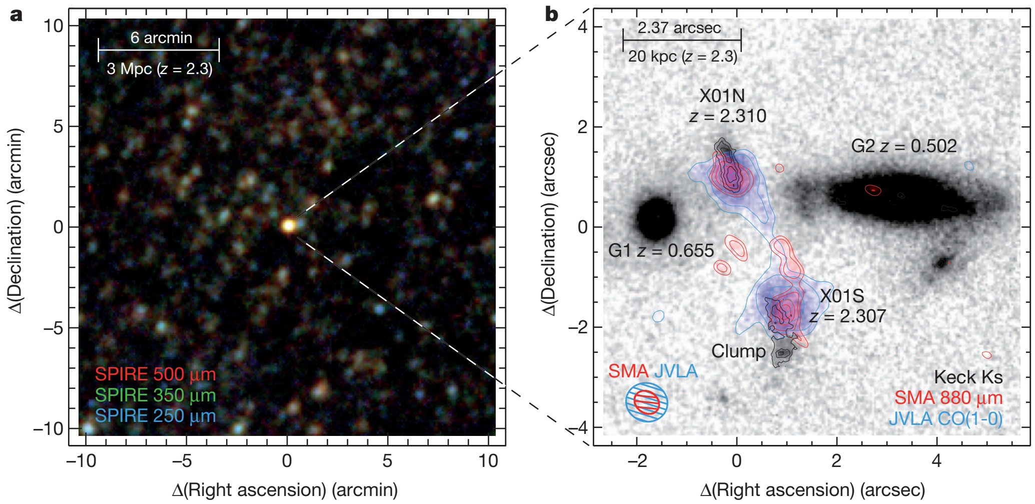


Figure 1 | Multi-wavelength view of HXMM01. a, The Herschel Space Observatory's false three-colour image combining 250-µm (blue), 350-µm (green) and 500-µm (red) images with resolutions of 18″, 25″ and 36″, respectively. HXMM01 is the brightest source in the image. **b**, The highest-resolution images of HXMM01. The background and black contours show the near-infrared camera (NIRC2 at Keck) K_S-band adaptive optics image. Overlaid are the dust continuum emission at 880 µm from the SMA (red contours) and the molecular CO $J = 1 \rightarrow 0$ emission from the JVLA (blue contours). The SMA and JVLA contours are drawn at +3, +4, +6 and +8 times the r.m.s. noise σ (where $\sigma = 0.67$ mJy per beam for SMA, and $\sigma = 19$ µJy beam^{-1} for JVLA), and the Keck contours are at $+5\sigma$, $+8\sigma$ and $+11\sigma$ (where $\sigma = 3.7 \times 10^{-3}$ µJy per pixel). The two major components of HXMM01 (X01N and X01S) and the foreground galaxies are labelled along with their redshifts. There is also a bridge of material detected at the $\gtrsim 5\sigma$ level between X01N and X01S in the SMA and JVLA images. We also label the southern part of X01S as a 'clump' because of its distinct optical spectral energy distribution compared to the rest of HXMM01. This clump could be a contaminating source, so we have excluded it in our stellar mass estimate of X01S. The ellipses at the lower left show the beam's full-width at half-maximum (0.54″ × 0.44″ for the SMA and 0.83″ × 0.77″ for the JVLA).

system that has accumulated a disproportionate amount of gas because of a higher accretion rate compared to the SFR in the past. Assuming a constant SFR with the observed value and no additional gas accretion, HXMM01 will reach a gas fraction of about 20% in just 70 Myr or so, making its stellar mass and gas fraction consistent with the simulated galaxies.

Because the components of HXMM01 are resolved in our Submillimetre Array (SMA, on Mauna Kea, Hawaii, USA) and Karl G. Jansky Very Large Array (JVLA, in New Mexico, USA) observations, we use them to directly measure the physical extent of the dust and gas emission. We find that the intrinsic sizes of the dusty star-forming regions ($A_{880} = 5$–$7\,\mathrm{kpc}^2$) are three to seven times smaller than the molecular gas reservoirs ($A_{\mathrm{CO}} = 15$–$50\,\mathrm{kpc}^2$). This is in direct conflict with the commonly used assumption that the gas and dust in SMGs have the same physical extent. This discrepancy has also been observed in several other SMGs[11]. When the sizes are used to measure gas and star-formation surface densities, we find that HXMM01 lies along the sequence of local and high-redshift starbursts in the Kennicutt–Schmidt relation[8] (see Supplementary Information). We also estimate the star-formation efficiency, ϵ_{SF}, the percentage of gas that is converted into stars on a dynamical timescale. We find that the two merging SMGs are remarkably efficient in forming stars with values of $\epsilon_{\mathrm{SF}} = (10 \pm 3)\%$ for X01N and $(41 \pm 10)\%$ for X01S. Similar to other starbursts[12], these efficiencies are about an order of magnitude higher

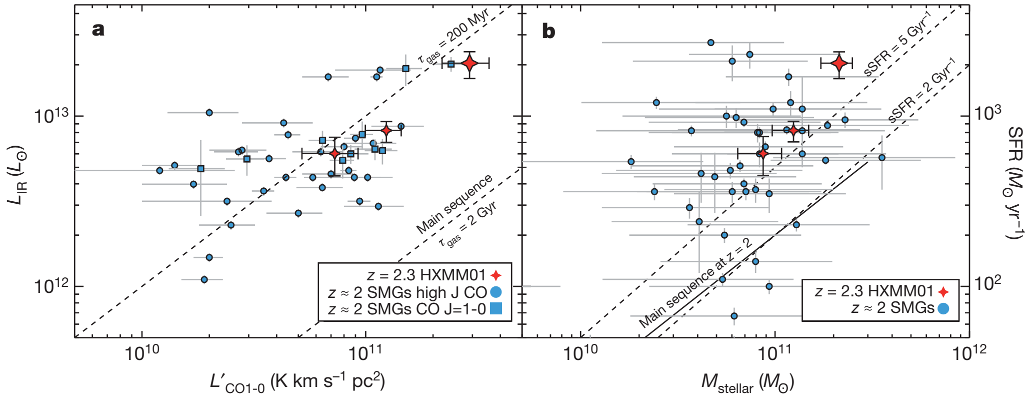


Figure 2 | The infrared luminosity and stellar masses of submillimetre galaxies. HXMM01 is plotted as the large filled red star. Its two main components, X01N and X01S, are plotted separately as the small filled red stars. **a**, Infrared luminosity L_{IR} plotted versus CO $J = 1 \rightarrow 0$ luminosity with blue squares showing SMGs with direct CO $J = 1 \rightarrow 0$ measurements[11,24]. The blue circles are SMGs with higher J CO line luminosities that have been converted to CO $J = 1 \rightarrow 0$ luminosities using the mean observed ratios[3]. Assuming $\alpha_{\mathrm{CO}} = 1$, the dashed lines indicate constant consumption timescales of the gas reservoir ($\tau_{\mathrm{gas}} \equiv 2M_{\mathrm{H2}}/\mathrm{SFR}$) of 200 Myr and 2 Gyr, which are the mean values for SMGs[3] and massive ($M_{\mathrm{star}} > 10^{11} M_\odot$) normal star-forming galaxies on the main sequence[25] at $z \approx 2$, respectively. **b**, SFR plotted versus stellar mass with blue data points showing SMGs at $z \approx 2$ using data from the literature[3,11,24,26]. The star-forming main sequence[27] at $z = 2$ is indicated by a solid line. Dashed lines show constant specific SFRs ($\mathrm{sSFR} \equiv \mathrm{SFR}/M_{\mathrm{star}}$). We find a sSFR for HXMM01 of $9.6 \pm 2.3\,\mathrm{Gyr}^{-1}$. At $z \approx 2$ this makes the value for HXMM01 five times higher than the main sequence of normal star-forming galaxies[28]. All error bars are formal $\pm 1\sigma$ standard deviations, including those associated with the lensing magnification. Thus HXMM01 has one of the largest gas and stellar content compared to the SMG population at $z \approx 2$, and it is clearly in a starburst phase.

Table 1 | Resolved mergers of submillimetre bright galaxies with projected separations of less than 30 kpc

Object	Redshift, z	Flux density at 850 µm, S_{850} (mJy)	Separation (kpc)	Molecular gas mass, M_{H2} ($10^{10} M_\odot$)	SFR ($M_\odot$ yr^{-1})
SMM J02399–0136	2.8	9	8	8.0	990
SMM J09431+4700	3.4	10.5	30	6.6	1,160
SMM J105141	1.2	8.7	5	6.6	960
SMM J123707/HDF 242	2.5	4.7	22	7.4	520
SMM J123711/HDF 254	2.0	4.2	8	0.8	650
SMM 163650/N2 850.4	2.4	8	4	9.6	880
4C 60.07	3.8	11	25	6.4	1,200
HXMM01	2.3	20 ± 4	19 ± 4	23 ± 6	2,000 ± 400

M_{H2} values are estimated using the conversion factor from CO $J = 1 \rightarrow 0$ luminosity to molecular gas mass, α_{CO}, in units of $M_\odot$/(K km s^{-1} pc^2). This factor is estimated to be about 4 for normal star-forming galaxies like the Milky Way[20] and about 1 for local starbursts[21]. Two independent methods outlined in the Supplementary Information, using either the gas-to-dust mass ratio or the velocity-integrated CO $J = 1 \rightarrow 0$ brightness temperature[22], lead to $\alpha_{CO} \approx 0.8$ for HXMM01, which is consistent with the α_{CO} measurements for several other SMGs[7,23]. Using this conversion factor, we find that HXMM01 contains $(2.3 \pm 0.6) \times 10^{11} M_\odot$ of molecular gas. We use the same $\alpha_{CO} = 0.8$ for other SMGs and have used CO $J = 1 \rightarrow 0$ measurements wherever available. SFRs are estimated using the 850-µm flux density to give SFR $\approx 110 S_{850}$ and have a 50% uncertainty. For HXMM01, this relation gives a SFR consistent with that from the total infrared luminosity. As for HXMM01, the values for SMM J02399–0136 have been corrected for lensing magnification.

than the typical value of normal star-forming galaxies ($\epsilon_{SF} \approx 2\%$). These efficiencies reach the theoretical maximum of 30% that has been estimated for molecular-cloud-dominated starbursts[13].

HXMM01 illustrates the rapid formation of an extremely massive elliptical galaxy in an equal-mass gas-rich merger. The high star-formation efficiency coupled with the short dynamical timescale will quickly turn this pair of starburst galaxies into a passive elliptical galaxy, because its SFR will decline with an e-folding time of 230 ± 40 Myr ($\tau_{gas} \equiv 2M_{H2}$/SFR) (where M_{H2} is the molecular gas mass) even without any feedback process[14] ejecting gas out of the galaxy. We have assumed no additional gas infall. The factor of two in our gas-exhausting timescale (τ_{gas}) accounts for the 50% gas recycling in stellar evolution. The descendant of HXMM01 is expected to have a stellar mass of about $4 \times 10^{11} M_\odot$, comparable to the most massive elliptical galaxies[4] at $z \approx 2$. Its stellar population will appear old and passively evolving by $z \approx 1.7$ (that is, 1 Gyr after the observed epoch) because the star formation will have ceased so rapidly[15]. Although gas-poor ("dry") merging[16] between less massive ellipticals is generally invoked to explain the subsequent stellar mass build-up after the SMG phase, the formation of a very massive galaxy directly through a gas-rich merger shows that dry mergers are not the only way to form the most massive ellipticals. They can also form by *in situ* starbursts involved in massive SMG mergers, similar to HXMM01.

It is extremely rare to see starburst mergers similar to HXMM01 because they are short-lived and unusually massive. In addition, only a small fraction of mergers are expected to be observed as two separate galaxies[17]. Therefore, although we estimate only one SMG–SMG merger as bright and luminous as HXMM01 to be present in every hundred square degrees or so, the importance of such mergers in galaxy evolution must be evaluated after correcting for their limited visibility, which is due to the wide separation of the merging galaxies and the short lifetime of the starburst phase. We can estimate the space density of massive mergers like HXMM01 from its CO luminosity, assuming all SMGs with such high CO luminosities are also mergers. But because the space density of galaxies as a function of CO luminosities is unknown, we resort to the observed space density of SMGs as a function of the SFR[18], given that SFR is proportional to gas mass. The space density of SMGs with SFR greater than 2,000 $M_\odot$ yr^{-1} is $(2.4 \pm 1.2) \times 10^{-6}$ Mpc^{-3} at $2 < z < 3$. Given that the time span covered by the redshift range is 1.2 Gyr and the lifetime of HXMM01 is about 200 Myr, the space density of HXMM01-like galaxies is $(1.4 \pm 0.7) \times 10^{-5}$ Mpc^{-3} after correcting for the 17% duty cycle. Despite the large uncertainty, this is comparable to the space density of passive elliptical galaxies[19] with stellar masses above $2 \times 10^{11} M_\odot$ at $z \approx 1.1$. Therefore, although SMGs as luminous as

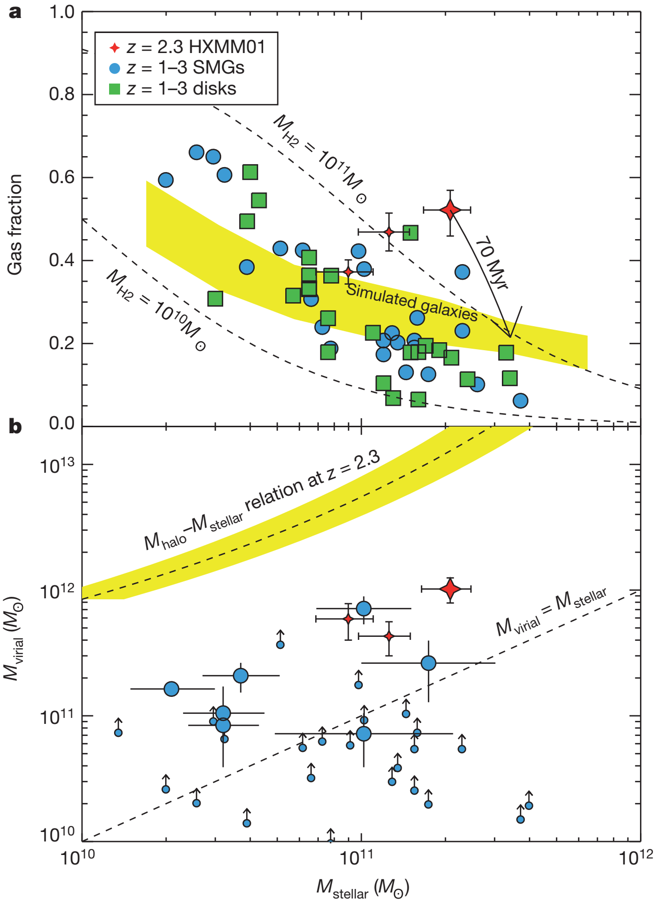


Figure 3 | The gas mass and dynamical state of submillimetre galaxies. **a**, The gas-to-baryon fraction versus stellar mass for star-forming galaxies. HXMM01 is plotted as the large filled red star, while its two main components, X01N and X01S, are the two small filled red stars. Because a significant fraction of CO $J = 1 \rightarrow 0$ emission is detected outside the two main components, the total gas fraction is higher than those of X01N and X01S. In **a**, the other data points are for SMGs (blue circles) and main-sequence star-forming galaxies (green squares). Both samples have used gas masses estimated from a model-calibrated relation between α_{CO} and CO brightness temperature[29]. The same relation is used to estimate α_{CO} for HXMM01. The yellow stripe shows $z \approx 2$ star-forming galaxies from cosmological hydrodynamic simulations[10]. The solid curve with an arrow shows the expected evolution of HXMM01 over the next 70 Myr, assuming the observed rate of star formation and the conservation of mass. The dashed curves indicate constant gas masses of $10^{10} M_\odot$ and $10^{11} M_\odot$. **b**, The virial dynamical mass versus stellar mass for SMGs at $z \approx 2$ with bigger and smaller blue circles showing SMGs from two other studies[3,17]. The yellow stripe shows the relation of stellar mass to halo mass at $z = 2.3$ from abundance matching[30]. All error bars are formal $\pm 1\sigma$ standard deviation, including those associated with the lensing magnifications. HXMM01 is extremely gas-rich compared to other $z \approx 2$ star-forming galaxies with similar stellar masses. It represents a galaxy evolution stage that is difficult to reproduce in simulations. The stellar mass of HXMM01 also implies a halo mass of about $5 \times 10^{12} M_\odot$ for each of the merging galaxies; that is, a merged halo mass of $10^{13} M_\odot$.

HXMM01 are rare, they could represent a short but critical transitional phase in the early formation of the most massive elliptical galaxies. This conclusion will soon be tested with wide-area submillimetre surveys that will uncover more SMG–SMG mergers similar to HXMM01.

Received 15 February; accepted 12 April 2013.

Published online 22 May; corrected online 19 June 2013 (see full-text HTML version for details).

1. McCarthy, P. J. *et al.* Evolved galaxies at z>1.5 from the Gemini Deep Deep Survey: the formation epoch of massive stellar systems. *Astrophys. J.* **614,** L9–L12 (2004).
2. Barger, A. J. *et al.* Submillimetre-wavelength detection of dusty star-forming galaxies at high redshift. *Nature* **394,** 248–251 (1998).
3. Bothwell, M. S. *et al.* A survey of molecular gas in luminous sub-millimetre galaxies. *Mon. Not. R. Astron. Soc.* **429,** 3047–3067 (2013).
4. Damjanov, I. *et al.* Red nuggets at z ~ 1.5: compact passive galaxies and the formation of the Kormendy Relation. *Astrophys. J.* **695,** 101–115 (2009).
5. Wardlow, J. L. *et al.* HerMES: candidate gravitationally lensed galaxies and lensing statistics at submillimeter wavelengths. *Astrophys. J.* **762,** 59–86 (2013).
6. Oliver, S. J. *et al.* The Herschel Multi-tiered Extragalactic Survey: HerMES. *Mon. Not. R. Astron. Soc.* **424,** 1614–1635 (2012).
7. Magnelli, B. *et al.* Dust temperature and CO→H_2 conversion factor variations in the SFR-M_* plane. *Astron. Astrophys.* **548,** A22 (2012).
8. Kennicutt, J. & Robert, C. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498,** 541–552 (1998).
9. da Cunha, E., Charlot, S. & Elbaz, D. A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. *Mon. Not. R. Astron. Soc.* **388,** 1595–1617 (2008).
10. Davé, R. *et al.* The nature of submillimetre galaxies in cosmological hydrodynamic simulations. *Mon. Not. R. Astron. Soc.* **404,** 1355–1368 (2010).
11. Ivison, R. J. *et al.* Tracing the molecular gas in distant submillimetre galaxies via CO(1–0) imaging with the Expanded Very Large Array. *Mon. Not. R. Astron. Soc.* **412,** 1913–1925 (2011).
12. Genzel, R. *et al.* A study of the gas-star formation relation over cosmic time. *Mon. Not. R. Astron. Soc.* **407,** 2091–2108 (2010).
13. Murray, N., Quataert, E. & Thompson, T. A. The disruption of giant molecular clouds by radiation pressure and the efficiency of star formation in galaxies. *Astrophys. J.* **709,** 191–209 (2010).
14. Di Matteo, T., Springel, V. & Hernquist, L. Energy input from quasars regulates the growth and activity of black holes and their host galaxies. *Nature* **433,** 604–607 (2005).
15. Martin, D. C. *et al.* The UV-optical galaxy color-magnitude diagram. III. Constraints on evolution from the blue to the red sequence. *Astrophys. J.* **173** (Suppl.), 342–356 (2007).
16. van Dokkum, P. G. The recent and continuing assembly of field elliptical galaxies by red mergers. *Astron. J.* **130,** 2647–2665 (2005).
17. Engel, H. *et al.* Most submillimeter galaxies are major mergers. *Astrophys. J.* **724,** 233–243 (2010).
18. Chapman, S. C., Blain, A. W., Smail, I. & Ivison, R. J. A redshift survey of the submillimeter galaxy population. *Astrophys. J.* **622,** 772–796 (2005).
19. Ilbert, O. *et al.* Galaxy stellar mass assembly between $0.2 < z < 2$ from the S-COSMOS survey. *Astrophys. J.* **709,** 644–663 (2010).
20. Strong, A. W. & Mattox, J. R. Gradient model analysis of EGRET diffuse galactic gamma-ray emission. *Astron. Astrophys.* **308,** L21–L24 (1996).
21. Downes, D. & Solomon, P. M. Rotating nuclear rings and extreme starbursts in ultraluminous galaxies. *Astrophys. J.* **507,** 615–654 (1998).
22. Narayanan, D., Krumholz, M. R., Ostriker, E. C. & Hernquist, L. A general model for the CO-H2 conversion factor in galaxies with applications to the star formation law. *Mon. Not. R. Astron. Soc.* **421,** 3127–3146 (2012).
23. Magdis, G. E. *et al.* GOODS-Herschel: gas-to-dust mass ratios and CO-to-H2 conversion factors in normal and starbursting galaxies at high-z. *Astrophys. J.* **740,** L15–L20 (2011).
24. Riechers, D. A., Hodge, J., Walter, F., Carilli, C. L. & Bertoldi, F. Extended cold molecular gas reservoirs in $z\approx3.4$ submillimeter galaxies. *Astrophys. J.* **739,** L31–L36 (2011).
25. Genzel, R. *et al.* The metallicity dependence of the CO → H_2 conversion factor in $z \geq 1$ star-forming galaxies. *Astrophys. J.* **746,** 69–79 (2012).
26. Hainline, L. J. Multi-Wavelength Properties of Submillimeter-Selected Galaxies. PhD thesis, *Cal. Inst. Technol*. (2008).
27. Daddi, E. *et al.* Multiwavelength study of massive galaxies at $z \sim 2$. I. Star formation and galaxy growth. *Astrophys. J.* **670,** 156–172 (2007).
28. Elbaz, D. *et al.* GOODS-Herschel: an infrared main sequence for star-forming galaxies. *Astron. Astrophys.* **533,** A119 (2011).
29. Narayanan, D., Bothwell, M. & Davé, R. Galaxy gas fractions at high redshift: the tension between observations and cosmological simulations. *Mon. Not. R. Astron. Soc.* **426,** 1178–1184 (2012).
30. Behroozi, P. S., Conroy, C. & Wechsler, R. H. A comprehensive analysis of uncertainties affecting the stellar mass-halo mass relation for $0 < z < 4$. *Astrophys. J.* **717,** 379–403 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements Herschel is an ESA space observatory with science instruments provided by European-led Principal Investigator consortia and with important participation from NASA. This research has made use of data from the HerMES project (http://hermes.sussex.ac.uk/). HerMES is a Herschel Key Programme using Guaranteed Time from the SPIRE instrument team, ESAC scientists and a mission scientist. SPIRE has been developed by a consortium of institutes led by Cardiff University (UK) and including the University of Lethbridge (Canada); NAOC (China); CEA, LAM (France); IFSI, the University of Padua (Italy); IAC (Spain); Stockholm Observatory (Sweden); Imperial College London, RAL, UCL-MSSL, UKATC, the University of Sussex (UK); and Caltech/JPL, IPAC and the University of Colorado (USA). This development has been supported by the following national funding agencies: CSA (Canada); NAOC (China); CEA, CNES and CNRS (France); ASI (Italy); MCINN (Spain); SNSB (Sweden); STFC (UK); and NASA (USA). The data presented in this paper will be released through the HeDaM Database in Marseille at http://hedam.oamp.fr/HerMES.

Author Contributions H.F. and A.C. wrote the manuscript and led the project. C.F., R.J.I., D.A.R., M.G., R.S.B. and A.I.H. contributed significantly to the taking and analysis of the follow-up data with various instruments. All other co-authors of this paper contributed extensively and equally by their varied contributions to the HerMES project, planning of HerMES observations, analysis of HerMES data, and by commenting on this manuscript as part of an internal review process.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.F. (haif@uci.edu) and A.C. (acooray@uci.edu).

doi:10.1038/nature12225

Volcanism on Mars controlled by early oxidation of the upper mantle

J. Tuff[1], J. Wade[1] & B. J. Wood[1]

Detailed information about the chemical composition and evolution of Mars has been derived principally from the SNC (shergottite–nakhlite–chassignite) meteorites, which are genetically related igneous rocks of Martian origin[1,2]. They are chemically and texturally similar to terrestrial basalts and cumulates, except that they have higher concentrations of iron and volatile elements such as phosphorus and chlorine and lower concentrations of nickel and other chalcophile (sulphur-loving) elements[3]. Most Martian meteorites have relatively young crystallization ages (1.4 billion years to 180 million years ago[4]) and are considered to be derived from young, lightly cratered volcanic regions, such as the Tharsis plateau[4,5]. Surface rocks from the Gusev crater analysed by the Spirit rover are much older (about 3.7 billion years old[6]) and exhibit marked compositional differences from the meteorites[7]. Although also basaltic in composition, the surface rocks are richer in nickel and sulphur and have lower manganese/iron ratios than the meteorites. This has led to doubts that Mars can be described adequately using the 'SNC model'. Here we show, however, that the differences between the compositions of meteorites and surface rocks can be explained by differences in the oxygen fugacity during melting of the same sulphur-rich mantle. This ties the sources of Martian meteorites to those of the surface rocks through an early (>3.7 billion years ago) oxidation of the uppermost mantle that had less influence on the deeper regions, which produce the more recent volcanic rocks.

That the SNC meteorites originated on Mars became apparent from their young crystallization ages[4] and fractionated rare-earth patterns[8,9], both indicating a planet-sized parent body, and most convincingly by the presence of trapped Martian atmosphere in shock-produced glass and basaltic matrix[10,11]. The compositions—spanning the range from basaltic, olivine-phyric volcanic rocks through pyroxene-cumulate nakhlites and orthopyroxenites to peridotitic chassignites—enable reconstruction of the composition of the silicate part of the parent body using similar techniques to those used to derive the composition of the silicate Earth[3,12]. The results (Table 1) point to a relatively oxidized Mars, with a higher FeO content of the mantle than Earth, and enrichments relative to Earth in moderately volatile elements such as Na, K, Mn, P and the halogens. Mars' putative mantle is not, however, enriched relative to Earth in volatile chalcophile elements, such as In and Tl; Dreibus and Wänke[3] estimated a significant depletion in Ni. They ascribed the latter trends to extraction of chalcophile elements from the mantle into a S-rich core.

Dreibus and Wänke[3] modelled volatile-rich Mars as being assembled from two components, namely component A (volatile-depleted, highly-reduced) and component B (with a composition similar to that of oxidized CI chondrite, and volatile-rich). They adjusted the mixing ratio to 65:35 (A:B) to fit Mars' density and a chondritic Fe/Si ratio, obtaining a core with about 14% S into which chalcophile elements were strongly partitioned. A core of about this composition together with the mantle composition of Table 1 are broadly consistent with more recent density and moment of inertia measurements[13–15]. The genetic relationships between the SNC meteorites and surface rocks need, therefore, to be viewed in the context of a sulphur-rich planet.

The Mars Exploration Rovers (MERs) Spirit and Opportunity performed APXS (alpha-proton X-ray spectrometry) measurements of chemical compositions of soil and rock near landing sites in the Gusev crater and Meridiani Planum, respectively. The rocks in the latter region are altered basaltic sandstones cemented by evaporitic salts, and are not genetically comparable to the igneous SNC meteorites[16]. The Gusev crater rocks are, however, largely basaltic in composition, showing similar major element, MgO, FeO, Al_2O_3, CaO and SiO_2 contents to the basaltic meteorites[17]. Although minor Cr is also present at similar concentrations, there are substantial differences in some minor elements from the corresponding SNCs. McSween *et al.*[7] show that the Gusev crater rocks are, at fixed MgO content, ~5 times richer in Ni than corresponding shergottites and are similar to terrestrial basalts in Ni content (Fig. 1). This is a conundrum, because the relatively high Ni contents of terrestrial rocks are believed to reflect very high pressures of core formation (~40 GPa; ref. 18), which favoured enhanced partitioning of Ni into the silicate mantle over the metallic core. Because Mars is only 11% of Earth's mass, it is difficult to envisage a mantle–core partitioning scenario at comparable pressures on this planet. McSween *et al.* also show that the surface rocks have lower Mn/Fe ratios than the SNCs, suggesting that Mars may be lower in relatively volatile Mn than previously believed. In contrast, the Gusev crater soils and rocks are extremely rich in volatile sulphur, typically containing about 2.5% S, with some analyses suggesting three times this concentration, consistent with a sulphur-rich planet. The S concentrations are much greater than those of the SNC meteorites (typically 1,000–2,000 p.p.m. in basaltic and ~400 p.p.m. in peridotitic types) and lead us to ask whether sulphur is the clue to the differences between Gusev crater rocks and SNC meteorites.

One important property of sulphur is its variable oxidation state. At the oxygen fugacities recorded by shergottites[19,20], S is present in silicate melts in the S^{2-} oxidation state, dissolving in concentrations, depending on FeO content of the melt, of ~1,500 p.p.m. (ref. 21).

Table 1 | Compositions

	Silicate Mars[3]	Silicate Earth[12]
MgO (%)	30.2	37.8
Al_2O_3 (%)	3.02	4.44
SiO_2 (%)	44.4	44.9
CaO (%)	2.45	3.54
TiO_2 (%)	0.14	0.2
FeO (%)	17.9	8.05
Na_2O (%)	0.5	0.36
P_2O_5 (%)	0.16	0.02
Cr_2O_3 (%)	0.76	0.38
MnO (%)	0.46	0.13
K (p.p.m.)	315	240
S (p.p.m.)	~2,500*	250
Ni (p.p.m.)	400; 1,800*	1,960
In (p.p.m.)	14	11
Cl (p.p.m.)	44	17
Br (p.p.b.)	165	50

All values for Mars from ref. 3, except where indicated.
* Our estimates (see text).

[1] Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK.

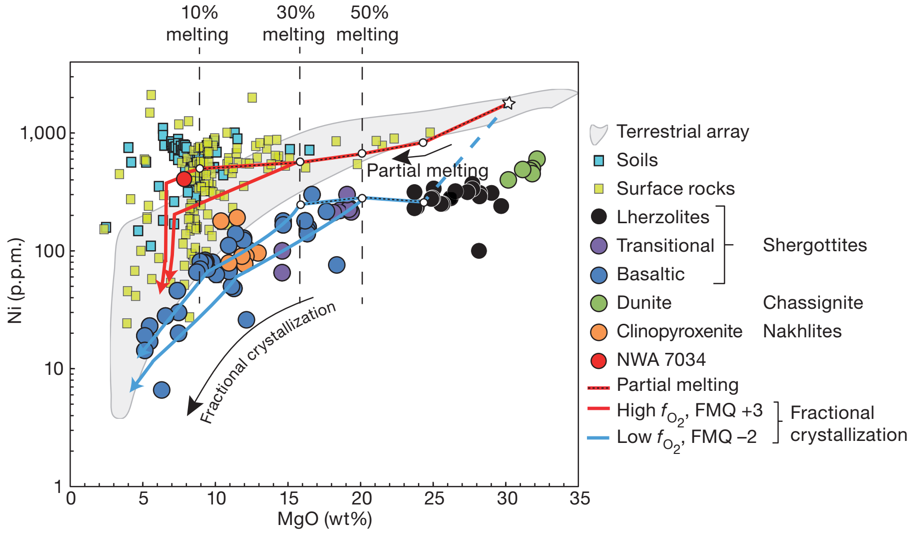


Figure 1 | MgO versus Ni contents of Martian meteorites and Martian surface materials. Meteorite data are from MetBase (http://www.metbase.de/). Surface materials are rocks and soils analysed by MER Spirit[16,28,29]. The terrestrial array of basalts, picrites and lherzolites (grey field) are from the GEOROC database (http://georoc.mpch-mainz.gwdg.de). Partial melting trends are from 1.5 GPa partial melts of ref. 3 Martian mantle[24], with initial Ni contents of melts calculated assuming batch-melting and a mantle Ni content of 1,800 p.p.m. Crystal fractionation trends were calculated using Petrolog[25] with Ni contents obtained by assuming Rayleigh fractionation (Methods). Ni–MgO trends of SNC meteorites (low f_{O_2}) and Gusev crater rocks (high f_{O_2}) can be generated from the same initial mantle composition depending only on the presence or absence of residual sulphide.

Terrestrial basalts, formed at oxygen fugacities below the fayalite–magnetite–quartz (FMQ) buffer are frequently saturated in sulphide, containing blobs of immiscible (Fe,Ni,Cu)S liquid and S concentrations of ~1,000 p.p.m. in the silicate melt[21,22]. At higher oxygen fugacities, S dissolves as S^{6+}, and sulphur becomes much more soluble in the silicate[23]. In terrestrial basalt compositions, for example, the S solubility at 2–3 log units above FMQ is ~1.5 wt% (ref. 23) and sulphate becomes stable. Given these large effects of oxygen fugacity on S solubility and the great differences in S content between Martian meteorites and surface rocks, we decided to explore the effects of oxygen fugacity on other elements of importance.

We start with the estimate of Martian mantle composition from ref. 3 and the experimentally produced partial melts of this mantle at 1.5 GPa (ref. 24). The partial melt compositions of ref. 3 at 10%, 30% and 50% degrees of melting[24] were allowed to fractionally crystallize the liquidus mantle minerals (olivine, orthopyroxene, clinopyroxene and spinel), using the crystallization program Petrolog[25] for major elements and the Rayleigh fractionation equation for the trace elements Ni and Mn (see Fig. 1 and below; see also Methods section). We also tested for the effects of sulphide saturation using the elements S, Ni and Fe. The extent of sulphide saturation is a function of melt composition, pressure, temperature and oxygen fugacity (f_{O_2}). We modelled the effect of sulphide under two conditions: 'sulphide present' (low f_{O_2}, about 2 log units below FMQ) and 'sulphide absent' (high f_{O_2}, 2–3 log units above FMQ). The low oxygen fugacity is consistent with oxygen fugacity values estimated for basaltic shergottites[20], and the high f_{O_2} values are consistent with the high concentrations of sulphur and sulphate in the surface rocks and with the presence of magnetite and maghemite in the Martian meteorite (NWA 7034) that most resembles the Gusev crater rocks[26].

As the mantle[3] melts, the residual phases are ~45% olivine and ~55% pyroxene, with minor spinel[27], which means that, because Ni is compatible (crystal–melt partition coefficients $D_{Ni}^{xtl/liq} > 1$) in the silicates, the melt must be lower in Ni than the mantle. It is therefore impossible to produce both Gusev crater rocks (Ni contents up to ~1,000 p.p.m.) and SNC meteorites from a source containing 400 p.p.m. Ni (Table 1). Mars' mantle must instead have approximately the same Ni content as the terrestrial mantle to explain the Gusev crater Ni–MgO trend of Fig. 1. The modelled curves on Fig. 1 were therefore produced with an assumed Ni content of the mantle of 1,800 p.p.m., appropriate values of the mineral–melt partition coefficients (Supplementary Information) and a range of $D_{Ni}^{sulph/sil}$ (partition coefficient for Ni between sulphide and silicate melt) values. As can be seen, the low Ni contents of the SNC meteorites are produced at appropriate low oxygen fugacity provided sufficient sulphide is present in the residual mantle. This would require, depending on $D_{Ni}^{sulph/sil}$, mantle S contents of 1,800–3,800 p.p.m. (Fig. 2), an order of magnitude greater than terrestrial mantle values, which could conceivably be due to less efficient extraction of accreting sulphide to the core on the smaller planet. Cumulates from fractional crystallization of 30–50% partial melts overlap the fields of lherzolitic and dunitic meteorites. At the higher values of f_{O_2} of the Gusev crater rocks, melting would consume all sulphur as SO_4^{2-} up to a saturation value of ~1.5% S (ref. 23), corresponding to 12–25% partial melting. At higher degrees of melting, no residual sulphide would remain in the mantle and Ni would be much more strongly partitioned into the melt, producing the

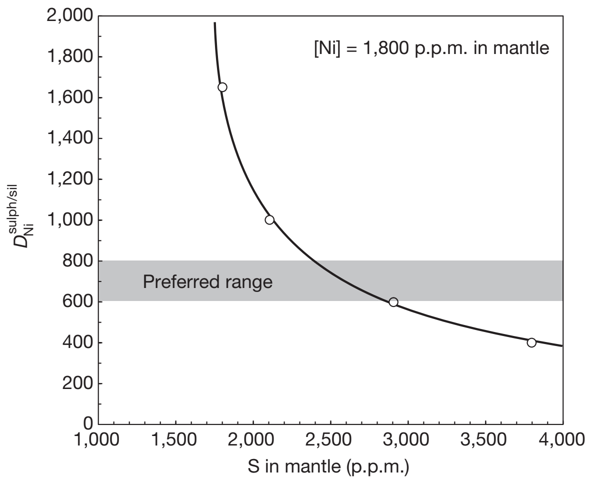


Figure 2 | Values of $D_{Ni}^{sulph/sil}$ and S contents of the Martian mantle used in our modelling. Here $D_{Ni}^{sulph/sil} = [Ni]_{sulphide}/[Ni]_{silicate}$. Values on this line generate the 'high f_{O_2}' and 'low f_{O_2}' curves of Fig. 1 at a Ni content of the mantle of 1,800 p.p.m. The preferred range of $D_{Ni}^{sulph/sil}$ values (shading) is based on experiments, and is discussed further in Methods.

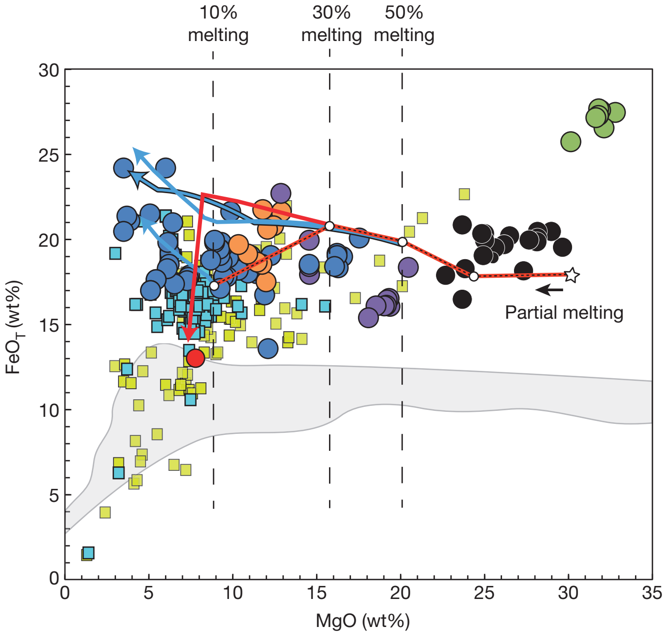


Figure 3 | Plot of FeO$_T$ versus MgO for initial partial melts of ref. 3 mantle[24] and fractionated liquids. The silicate liquids were fractionated at FMQ − 2 log units (low f_{O_2}) and FMQ + 3 log units (high f_{O_2}) using Petrolog[25]. Key as for Fig. 1. Note that the SNC meteorites show slight iron enrichment with decreasing MgO, indicative of low f_{O_2}, whereas at high f_{O_2} liquids decrease rapidly in FeO content once magnetite starts precipitating.

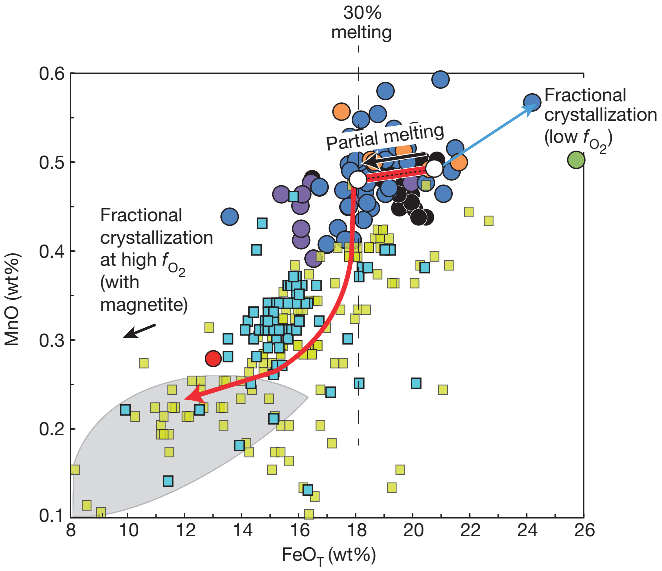


Figure 4 | Trends of MnO versus FeO for partial melts (30%) of the ref. 3 mantle and their derivative fractionally crystallizing liquids. Key as for Fig. 1. Note that the low f_{O_2} trend of fractional crystallization is towards high FeO, high MnO, whereas at high f_{O_2} FeO decreases as Fe_3O_4 precipitates and MnO follows it, because $D_{\mathrm{Mn}}^{\mathrm{Fe_3O_4/liq}} > 1$ (Supplementary Table 1).

Gusev crater trend of Fig. 1. Melting at higher oxygen fugacity also provides a first-order explanation for the high S contents of the surface rocks, although the very high sulphate contents[16,28] of some analyses suggests additional accumulation through hydrothermal processes[29].

In the later stages of crystallization at high f_{O_2} (2–3 log units above FMQ), magnetite begins precipitating. This reduces the Fe contents of the fractionating melts (Fig. 3), generating—because Mn and Ni are compatible in magnetite—decreasing Mn and Ni contents of the residual liquids (Figs 1 and 4). Thus, the proposed difference in oxygen fugacity between Gusev crater and SNC rocks also explains the differences in Mn–Fe trends and the presence of magnetite and maghemite in the 'Gusev crater-like' meteorite, NWA 7034. The SNC meteorites did not precipitate Fe_3O_4, containing instead Ti-rich members of the titanomagnetite series, indicative of low oxygen fugacity.

Our results demonstrate that the surface basaltic rocks of the Gusev crater and the igneous SNC meteorites are consistent with partial melting and fractional crystallization from the same source (similar in composition to that given in ref. 3) but under different f_{O_2} conditions. The observation that the generally older (~3.7 Gyr) Gusev crater rocks require more oxidized conditions than the younger meteorites leads to questions of the origin of the oxidation and the distribution of oxidized and reduced mantle regions. On Earth, some oxidation of the mantle occurs at subduction zones where surficial materials are recycled into the mantle. Above subduction zones, arc volcanics typically record oxygen fugacities several log units above those of mid-ocean-ridge basalt[30] and, in some cases, may, like the Gusev crater rocks, crystallize sulphate[31]. The presence of subduction on Mars is debated, but there have been suggestions that crust formation involved early plate tectonics[32], and mixing of oxidized near-surface rocks with reduced melts has been invoked to explain variations in f_{O_2} of shergottites[19,20].

The implications of our results are that Mars' surface oxidized early in its history, and that oxidized material was recycled into the upper mantle. This recycling may, we consider, be restricted to the shallow depths (~120 km) where the Gusev crater rocks originated by partial melting. The lower Al_2O_3 and Na_2O contents (Supplementary Information) and fractionated rare-earth element patterns of the SNC meteorites are consistent with a deeper reduced source region in the garnet field (>2.5 GPa) at depths >190 km that has not been so strongly affected by recycling of oxidized near-surface rocks.

METHODS SUMMARY

We took major element compositions of partial melts (10%, 30%, 50% melting) of mantle[3] from experiments at 1.5 GPa (ref. 24) and calculated fractional crystallization trends using Petrolog[25] software. Concentrations of trace elements Ni and Mn in the batch melts were obtained from the standard batch melting equation, assuming modal melting and including sulphide as a potential residual phase. Concentrations of Ni and Mn in the fractionally crystallized liquids were calculated by assuming Rayleigh fractionation.

The calculations were performed under two sets of conditions: at low f_{O_2} and at high f_{O_2}. Calculations at low f_{O_2} assumed sulphide saturation, which means that Ni was substantially partitioned into residual sulphide during batch melting and to some extent during fractional crystallization at sulphide saturation. The effects of varying the S content of the Martian mantle and corresponding $D_{\mathrm{Ni}}^{\mathrm{sulph/sil}}$ values required to fit the SNC trend are shown in Fig. 2. Calculations at high f_{O_2} assumed a solubility limit of S of 1.5% (ref. 23), which means that all S entered the melt except at the lowest degrees of partial melting. This increasing sulphur solubility augments the Ni contents of the product melts, relative to the low f_{O_2} case as shown in Fig. 1.

Magnetite was allowed to be a fractionating phase and was found (from Petrolog) to stabilize on the liquidus at high f_{O_2}. Appearance of magnetite reduces the Fe contents of residual melts (Fig. 3) and, assuming Rayleigh fractionation, their Mn and Ni contents (Fig. 4).

Full Methods and any associated references are available in the online version of the paper.

Received 22 February; accepted 18 April 2013.

1. McSween, H. Y. *et al.* Petrogenetic relationship between Allan Hills 77005 and other achondrites. *Earth Planet. Sci. Lett.* **45,** 275–284 (1979).
2. Walker, D., Stolper, E. M. & Hays, J. F. Basaltic volcanism: the importance of planet size. *Proc Lunar Planet. Sci. Conf.* **10,** 1995–2015 (1979).
3. Dreibus, G. & Wanke, H. Mars, a volatile-rich planet. *Meteoritics* **20,** 367–381 (1985).
4. Nyquist, L. E. *et al.* Ages and geologic histories of Martian meteorites. *Space Sci. Rev.* **96,** 105–164 (2001).
5. Treiman, A. H. The nakhlite meteorites: augite-rich igneous rocks from Mars. *Chemie Erde Geochem.* **65,** 203–270 (2005).
6. Greeley, R. *et al.* Fluid lava flows in Gusev crater, Mars. *J. Geophys. Res. Planets* **110,** E05008 (2005).
7. McSween, H. Y., Taylor, G. J. & Wyatt, M. B. Elemental composition of the martian crust. *Science* **324,** 736–739 (2009).

8. McSween, H. Y. SNC meteorites — are they martian rocks? *Geology* **12,** 3–6 (1984).
9. Shih, C. Y. *et al.* Chronology and petrogenesis of young achondrites, Shergotty, Zagami, and Alha77005 — late magmatism on a geologically active planet. *Geochim. Cosmochim. Acta* **46,** 2323–2344 (1982).
10. Becker, R. H. & Pepin, R. O. The case for a martian origin of the shergottites: nitrogen and noble gases in EETA-79001. *Earth Planet. Sci. Lett.* **69,** 225–242 (1984).
11. Bogard, D. D. & Johnson, P. Martian gases in an Antarctic meteorite. *Science* **221,** 651–654 (1983).
12. McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120,** 223–253 (1995).
13. Bertka, C. M. & Fei, Y. W. Density profile of an SNC model Martian interior and the moment-of-inertia factor of Mars. *Earth Planet. Sci. Lett.* **157,** 79–88 (1998).
14. Konopliv, A. S. *et al.* Mars high resolution gravity fields from MRO, Mars seasonal gravity, and other dynamical parameters. *Icarus* **211,** 401–428 (2011).
15. Rivoldini, A., Van Hoolst, T., Verhoeven, O., Mocquet, A. & Dehant, V. Geodesy constraints on the interior structure and composition of Mars. *Icarus* **213,** 451–472 (2011).
16. Rieder, R. *et al.* Chemistry of rocks and soils at Meridiani Planum from the alpha particle X-ray spectrometer. *Science* **306,** 1746–1749 (2004).
17. McSween, H. Y. *et al.* Characterization and petrologic interpretation of olivine-rich basalts at Gusev Crater, Mars. *J. Geophys. Res.* **111,** E02S10 (2006).
18. Wood, B. J., Walter, M. J. & Wade, J. Accretion of the Earth and segregation of its core. *Nature* **441,** 825–833 (2006).
19. Wadhwa, M. Redox state of Mars' upper mantle and crust from Eu anomalies in shergottite pyroxenes. *Science* **291,** 1527–1530 (2001).
20. Herd, C. D. K., Borg, L. E., Jones, J. H. & Papike, J. J. Oxygen fugacity and geochemical variations in the martian basalts: implications for martian basalt petrogenesis and the oxidation state of the upper mantle of Mars. *Geochim. Cosmochim. Acta* **66,** 2025–2036 (2002).
21. O'Neill, H. S. & Mavrogenes, J. A. The sulfide capacity and the sulfur content at sulfide saturation of silicate melts at 1400°C and 1 bar. *J. Petrol.* **43,** 1049–1087 (2002).
22. Mathez, E. A. Sulfur solubility and magmatic sulfides in submarine basalt glass. *J. Geophys. Res.* **81,** 4269–4276 (1976).
23. Jugo, P. J. Sulfur content at sulfide saturation in oxidized magmas. *Geology* **37,** 415–418 (2009).
24. Bertka, C. M. & Holloway, J. R. Anhydrous partial melting of an iron-rich mantle. 2. Primary melt compositions at 15 kbar. *Contrib. Mineral. Petrol.* **115,** 323–338 (1994).
25. Danyushevsky, L. V. & Plechov, P. Petrolog3: integrated software for modeling crystallization processes. *Geochem. Geophys. Geosyst.* **12,** Q07021 (2011).
26. Agee, C. B. *et al.* Unique meteorite from Early Amazonian Mars: water-rich basaltic breccia Northwest Africa 7034. *Science* **339,** 780–785 (2013).
27. Bertka, C. M. & Holloway, J. R. Anhydrous partial melting of an iron-rich mantle. 1. Subsolidus phase assemblages and partial melting phase-relations at 10 to 30 kbar. *Contrib. Mineral. Petrol.* **115,** 313–322 (1994).
28. Gellert, R. *et al.* Alpha particle X-ray spectrometer (APXS): results from Gusev crater and calibration report. *J. Geophys. Res.* **111,** E02S05 (2006).
29. Ming, D. W. *et al.* Geochemical and mineralogical indicators for aqueous processes in the Columbia Hills of Gusev crater, Mars. *J. Geophys. Res.* **111,** E02S12 (2006).
30. Carmichael, I. S. E. The redox states of basic and silicic magmas — a reflection of their source regions. *Contrib. Mineral. Petrol.* **106,** 129–141 (1991).
31. Luhr, J. F., Carmichael, I. S. E. & Varekamp, J. C. The 1982 eruptions of El Chichon Volcano, Chiapas, Mexico: mineralogy and petrology of the anhydrite-bearing pumices. *J. Volcanol. Geotherm. Res.* **23,** 69–108 (1984).
32. McSween, H. Y., Grove, T. L. & Wyatt, M. B. Constraints on the composition and petrogenesis of the Martian crust. *J. Geophys. Res.* **108,** 5135 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge support from the UK Science and Technology Facilities Council (grant ST/G00272X/1) and from the European Research Council (grant 267764).

Author Contributions The idea of this paper was developed by all three authors during an informal discussion. J.T. did most of the modelling with input from J.W. and B.J.W. B.J.W. wrote most of the paper with input from J.T. and J.W.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.J.W. (berniew@earth.ox.ac.uk).

METHODS

Our model consists of two parts: (1) partial melting based on the 1.5 GPa melting experiments of ref. 24, in which Martian mantle of composition given in ref. 3 was melted; (2) crystal fractionation of the partial melts of part (1), using the Petrolog fractionation program[25]. An annotated Excel spreadsheet, with examples, is provided as Supplementary Information.

The study of ref. 24 provided the major element phase compositions, degree of melting and pressures and temperatures for our calculations. The experimental results provide data at ~70%, 50%, 30% and 10% melting of the mantle of ref. 3 composition. Modal abundances of experimentally derived phases (olivine, orthopyroxene, clinopyroxene and spinel) were estimated from the subsolidus modal assemblage of ref. 27 and mass balance calculations[33]. Major element compositions during batch partial melting were taken from the experimental results[27,34].

For concentrations of trace elements Ni and Mn in the batch partial melts we used the batch melting equation:

$$\frac{C_L}{C_O}=\frac{1}{D+F(1-P)} \tag{1}$$

where C_L is the concentration of the trace element in the liquid, C_O is the original concentration in the mantle before the onset of melting, D is the bulk distribution coefficient (defined as $D=X_\alpha K_\alpha+X_\beta K_\beta+\ldots$, with X_α and X_β the weight fraction of phases α, β and so on, and K_α and K_β the solid–melt partition coefficients of phases α and β, respectively), F is the melt fraction and P represents the contribution of the various phases to the melt and is defined in a similar way to D. We assumed modal melting which gives $P=D$. The melt fractions, modal abundances of phases and temperature of the loss of each phase on melting were taken from partial melting experiments on the ref. 3 composition at 1.5 GPa (ref. 24). Assumed solid–melt partition coefficients K_{Ni} were 5 for olivine–melt and spinel–melt, 3 for orthopyroxene–melt and 2 for clinopyroxene–melt (data sources in Supplementary Table 1). Partition coefficients K_{Mn} were 0.9 for olivine–melt and orthopyroxene–melt, 1.0 for clinopyroxene–melt and 0 for spinel–melt (Supplementary Table 1). We derive $D_{Ni}^{sulph/sil}$ indirectly from olivine–sulphide liquid partitioning or directly from sulphide liquid–silicate liquid partitioning experiments. In the former case we use the observation[35] that, at low f_{O_2}, $\frac{[Fe/Ni]_{ol}}{[Fe/Ni]_{sulph}}$ is ~30. Then, given $D_{Fe}^{ol/liq}$ of about 1 for silicate liquid near the ref. 3 solidus[27] and $D_{Ni}^{ol/liq}$ of about 5 (Supplementary Table 1), $\frac{[Fe/Ni]_{sil}}{[Fe/Ni]_{sulph}}$ should be ~150. Taking sulphide liquid with approximately 60% Fe and a 10% batch silicate partial melt with 13.5% Fe[27] we obtain $D_{Ni}^{sulph/sil}$ of 670. This is in the middle of the range of values determined directly[36].

For our calculations at low f_{O_2} we assumed that the partial melts were sulphide saturated, which from the observed speciation of S in silicate melts[37], means that at 2 log units below FMQ the ratio of S^{6+}/S^{2-} in the melt is about 10^{-6}. Sulphide melt–silicate melt partition coefficients were taken to be 0.48 for Mn and 400–1,650 for Ni (Fig. 2, Supplementary Table 1). Given no S^{6+} the S content of the melt can be calculated directly from the sulphide content at sulphide saturation (SCSS)[21]

$$\ln\,[S]_{SCSS}=\frac{\Delta G^o}{RT}+\ln C_S+\ln\alpha_{FeS}^{sulphide}-\ln\alpha_{FeO}^{silicate} \tag{2}$$

where $[S]_{SCSS}$ is the S content of the silicate melt in p.p.m., C_S is the sulphide capacity of the melt, $\alpha_{FeS}^{sulphide}=1$ and $\alpha_{FeO}^{silicate}$ is the mole fraction of FeO in the silicate melt. ΔG° is the free energy of the equilibrium between silicate melt and an immiscible iron sulphide phase:

$$FeO+{}^1\!/_2\,S_2=FeS+{}^1\!/_2\,O_2 \tag{3}$$

From ref. 21 we have:

$$\Delta G^o=122{,}175-80.280T+8.474T\ln T \tag{4}$$

The sulphide capacity, C_s, was calculated using equation (20) in ref. 21:

$$\ln C_S=A_0+A_{Ca}X_{Ca}+A_{Mg}X_{Mg}+A_{Al}X_{Al}+A_{Na/K}(X_{Na}+X_K)+A_{Ti}X_{Ti}+A_{Fe}X_{Fe}\pm B_{Fe\text{-}Ti}X_{Fe}X_{Ti} \tag{5}$$

with fit parameters from their table 12 and the melt compositions in this study.

From the calculated S content of the liquid during batch partial melting at low f_{O_2} we calculate, for a given S content of the mantle, the fraction of residual sulphide. When combined with the proportions of silicate phases[24] this enables us to calculate the Ni and Mn contents of the partial melts from equation (1).

As oxygen fugacity increases, sulphide becomes more soluble in silicate melt as the fraction of S^{6+} increases[37]:

$$[Total\_S]=[S]_{SCSS}[1+10^{(2\Delta FMQ-2.1)}] \tag{6}$$

In equation (6) ΔFMQ refers to the oxygen fugacity in log units relative to the FMQ buffer. Thus, at ~1.5 log units above FMQ the total S content of the melt at sulphide saturation is 10 times that at low f_{O_2} and approaches the value of ~1.5% S at sulphate saturation[23]. For our calculations at high f_{O_2} (3 log units above FMQ) we therefore assumed that, during batch partial melting, S dissolves in the melt until the mantle content of S is exhausted or 15,000 p.p.m. S in the melt is reached. Thus, sulphide would not be present in the residue except for very low degrees of partial melting. Ni and Mn contents of the product batch melts were then calculated, as before using equation (1).

Batch partial melts (10, 30 and 50% melt) were allowed to 'cool' and fractionally crystallize olivine, orthopyroxene, clinopyroxene and spinel using Petrolog[25] with the extent of sulphide precipitation controlled by S solubility in the melt. During fractional crystallization at low f_{O_2} the Mg contents of the melts decrease, leading to increasing solubility of S and suppression of sulphide precipitation. The result is that Ni is retained in the melt and less strongly partitioned into the cumulates than would otherwise be expected. Cumulate compositions overlap the fields of lherzolitic shergottites and dunitic chassignite in Fig. 1.

Magnetite, which is observed in meteorite NWA 7034, was also included as a potential crystallizing phase within the Petrolog calculations. These were performed under 'high f_{O_2}' (that is, 3 log units above the FMQ buffer) and 'low f_{O_2}' (that is, 2 log units below the FMQ buffer). In the fractionation calculations magnetite was present under high f_{O_2} but absent under low f_{O_2} conditions.

Rayleigh fractionation was used to calculate the behaviour of Ni and Mn from the 10, 30 and 50% experimentally determined partial melts:

$$\frac{C_L}{C_O}=F'^{(D-1)} \tag{7}$$

where C_L and C_O are the melt composition and initial composition, respectively, F' is the fraction of melt remaining at each step, and D, the bulk distribution coefficient, is defined as before.

33. Walter, M. J. Melting of garnet peridotite and the origin of komatiite and depleted lithosphere. *J. Petrol.* **39,** 29–60 (1998).
34. Agee, C. B. & Draper, D. S. Experimental constraints on the origin of Martian meteorites and the composition of the Martian mantle. *Earth Planet. Sci. Lett.* **224,** 415–429 (2004).
35. Brenan, J. M. Effects of fO$_2$, fS$_2$, temperature and melt composition on Fe-Ni exchange between olivine and sulfide liquid: implications for natural olivine-sulfide assemblages. *Geochim. Cosmochim. Acta* **67,** 2663–2681 (2003).
36. Li, Y. & Audetat, A. Partitioning of V, Mn, Co, Ni, Cu, Zn, As, Mo, Ag, Sn, Sb, W, Au, Pb, and Bi between sulfide phases and hydrous basanite melt at upper mantle conditions. *Earth Planet. Sci. Lett.* **355–356,** 327–340 (2012).
37. Jugo, P. J., Wilke, M. & Botcharnikov, R. E. Sulfur K-edge XANES analysis of natural and synthetic basaltic glasses: implications for S speciation and S content as function of oxygen fugacity. *Geochim. Cosmochim. Acta* **74,** 5926–5938 (2010).

doi:10.1038/nature12226

Masses of exotic calcium isotopes pin down nuclear forces

F. Wienholtz[1], D. Beck[2], K. Blaum[3], Ch. Borgmann[3], M. Breitenfeldt[4], R. B. Cakirli[3,5], S. George[1], F. Herfurth[2], J. D. Holt[6,7], M. Kowalska[8], S. Kreim[3,8], D. Lunney[9], V. Manea[9], J. Menéndez[6,7], D. Neidherr[2], M. Rosenbusch[1], L. Schweikhard[1], A. Schwenk[7,6], J. Simonis[6,7], J. Stanja[10], R. N. Wolf[1] & K. Zuber[10]

The properties of exotic nuclei on the verge of existence play a fundamental part in our understanding of nuclear interactions[1]. Exceedingly neutron-rich nuclei become sensitive to new aspects of nuclear forces[2]. Calcium, with its doubly magic isotopes ^{40}Ca and ^{48}Ca, is an ideal test for nuclear shell evolution, from the valley of stability to the limits of existence. With a closed proton shell, the calcium isotopes mark the frontier for calculations with three-nucleon forces from chiral effective field theory[3–6]. Whereas predictions for the masses of ^{51}Ca and ^{52}Ca have been validated by direct measurements[4], it is an open question as to how nuclear masses evolve for heavier calcium isotopes. Here we report the mass determination of the exotic calcium isotopes ^{53}Ca and ^{54}Ca, using the multi-reflection time-of-flight mass spectrometer[7] of ISOLTRAP at CERN. The measured masses unambiguously establish a prominent shell closure at neutron number $N = 32$, in excellent agreement with our theoretical calculations. These results increase our understanding of neutron-rich matter and pin down the subtle components of nuclear forces that are at the forefront of theoretical developments constrained by quantum chromodynamics[8].

Exotic nuclei with extreme neutron-to-proton asymmetries exhibit shell structures generated by unexpected orderings of shell occupations. Their description poses enormous challenges, because most theoretical models have been developed for nuclei at the valley of stability. It is thus an open question how well they can predict new magic numbers emerging far from stability[9–11]. This is closely linked to our understanding of the different components of the strong force between neutrons and protons, such as the spin–orbit or tensor interactions, which modify the gaps between single-particle orbits[12], and of three-body forces, which are pivotal in calculations of extreme neutron-rich systems based on nuclear forces[2,13,14]. The resulting magic numbers, as well as the strength of the corresponding shell closures, are critical for global predictions of the nuclear landscape[15], and thus for the successful modelling of matter in astrophysical environments.

Three-body forces arise naturally in chiral effective field theory[8], which provides a systematic basis for nuclear forces connected via its symmetries to the underlying theory of quarks and gluons, namely quantum chromodynamics. Owing to the consistent description in effective field theory, there are only two undetermined low-energy couplings in chiral three-nucleon forces at leading and sub-leading orders. These are constrained by the properties of light nuclei ^{3}H and ^{4}He only, so that all heavier elements are predictions in chiral effective field theory. The present frontier of three-nucleon forces is located in the calcium isotopes, where the structural evolution is dominated by valence neutrons due to the closed proton shell at atomic number $Z = 20$ (refs 3, 5). These predictions withstood a recent challenge from direct Penning-trap mass measurements of ^{51}Ca and ^{52}Ca at TITAN/TRIUMF[4], which have established a substantial change from the previous mass evaluation and leave completely open how nuclear masses evolve past ^{52}Ca. This region is also very exciting because of evidence of a new magic neutron number $N = 32$ from nuclear spectroscopy[16–18], with a high 2^+ excitation energy in ^{52}Ca (refs 19, 20). These results are accompanied by successful theoretical studies based on phenomenological shell-model interactions[21,22], which are similar for the excitation spectra at $N = 32$ but disagree markedly in their predictions for ^{54}Ca and further away from stability.

Here we present the first mass measurements of the exotic calcium isotopes ^{53}Ca and ^{54}Ca. These provide key masses for all theoretical models, and unambiguously establish a strong shell closure, in excellent agreement with the predictions including three-nucleon forces.

The mass of a nucleus provides direct access to the binding energy, the net result of all interactions between nucleons. Penning traps have proven to be the method of choice when it comes to high-precision mass determination of exotic nuclei[23,24]. The mass m of an ion of interest with charge q stored in a magnetic field B is determined by comparing its cyclotron frequency $\nu_C = qB/(2\pi m)$ to that of a well-known reference ion, $\nu_{C,Ref}$. The frequency ratio $r_{ICR} = \nu_{C,Ref}/\nu_C$ (ICR, ion cyclotron resonance) then yields the mass ratio directly and thus the atomic mass of the isotope.

We have made a critical step towards determining the pivotal calcium masses by introducing a new method of precision mass spectrometry for short-lived isotopes. The developments and measurements were performed with ISOLTRAP[25], a high-resolution Penning-trap mass spectrometer at the ISOLDE/CERN facility. This method was used to confirm and even improve the accuracy of the recent mass measurements

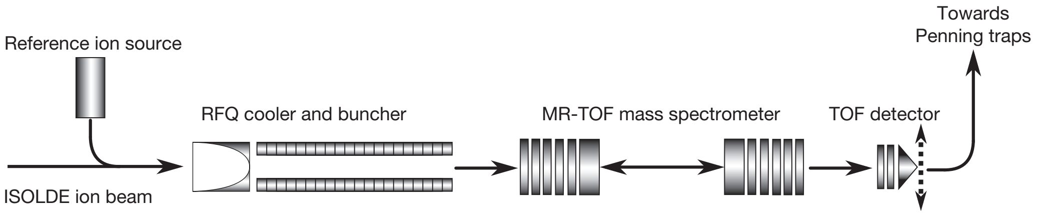


Figure 1 | Experimental set-up. Main components relevant for the ^{53,54}Ca study: incoming ISOLDE ion beam, reference ion source, radio-frequency quadrupole (RFQ) buncher, multi-reflection time-of-flight (MR-TOF) mass spectrometer and (removable) time-of-flight ion detector.

[1] Ernst-Moritz-Arndt-Universität Greifswald, Institut für Physik, Felix-Hausdorff-Strasse 6, D-17489 Greifswald, Germany. [2] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstrasse 1, D-64291 Darmstadt, Germany. [3] Max-Planck-Institut für Kernphysik, Saupfercheckweg 1, D-69117 Heidelberg, Germany. [4] Instituut voor Kern- en Stralingsfysica, Katholieke Universiteit, Celestijnenlaan 200d – bus 2418, B-3001 Heverlee, Belgium. [5] University of Istanbul, Department of Physics, 34134 Istanbul, Turkey. [6] Institut für Kernphysik, Technische Universität Darmstadt, D-64289 Darmstadt, Germany. [7] ExtreMe Matter Institute EMMI, GSI Helmholtzzentrum für Schwerionenforschung GmbH, D-64291 Darmstadt, Germany. [8] CERN, Geneva 23, CH-1211 Geneva, Switzerland. [9] CSNSM-IN2P3-CNRS, Université Paris-Sud, 91405 Orsay, France. [10] Institut für Kern- und Teilchenphysik, Technische Universität Dresden, Zellescher Weg 19, D-01069 Dresden, Germany.

of ^{51}Ca and ^{52}Ca (ref. 4). To advance past ^{52}Ca, we added a multi-reflection time-of-flight mass spectrometer/separator[26] (MR-TOF MS, see Fig. 1) to the three other ion traps that constitute ISOLTRAP, namely a linear Paul trap and two Penning traps (the latter are not shown in Fig. 1). In the MR-TOF MS, flight paths of several kilometres are folded into table-top dimensions. This device provides not only a mass-resolving power of more than 10^5, but also a mass uncertainty in the sub-parts-per-million (sub-p.p.m.) range. As typical flight times are about 10 ms, nuclides with half-lives of the same order are accessible. Likewise, nuclei with a lower production rate can be accessed, pushing the limits currently set by Penning-trap mass spectrometry to isotopes farther away from stability.

The neutron-rich calcium isotopes were produced at the online isotope separator ISOLDE in proton-induced fission reactions of a uranium carbide target at 1.4 GeV proton energy. The nuclides of interest were ionized by a highly selective, three-step laser-excitation scheme[27]. The ions were accelerated and transported to the ISOLTRAP set-up via ISOLDE's high-resolution separator as an essentially continuous 30 keV beam. They were captured and cooled in the radio-frequency quadrupole (RFQ) buncher and forwarded to the MR-TOF MS as bunches of about 60 ns duration. In the case of ^{51}Ca$^+$ and ^{52}Ca$^+$, the MR-TOF MS was operated as an isobar separator, delivering the purified bunches to the Penning traps, where the mass measurements were performed by determining the cyclotron-frequency ratios as described above. Nevertheless, for ^{53}Ca$^+$ and ^{54}Ca$^+$ the Penning-trap measurements were not possible because of the low production rates and copious isobaric contamination. For example, only a few ^{54}Ca ions per minute were detected behind the MR-TOF system, accompanied by several thousand contaminating ^{54}Cr ions. The rate of delivery of ^{54}Ca$^+$ to the Penning traps was considerably reduced owing to the lower transport efficiency and the decay losses caused by the required extra ion trapping time.

Thus for ^{53}Ca$^+$ and ^{54}Ca$^+$ the MR-TOF device itself was employed as a mass spectrometer, where the time of flight t of an ion is related to the mass-over-charge ratio m/q by $t = \alpha(m/q)^{1/2} + \beta$. Measuring the time of flight of two well-known reference ions, here ^{39}K and $^{53/54}$Cr (see Fig. 2), determines the experimental parameters α and β. With this calibration the mass m of the ions of interest, ^{53}Ca and ^{54}Ca, results directly from their time of flight. This relation can be expressed by $m^{1/2} = C_{\mathrm{TOF}}\Delta_{\mathrm{Ref}} + \Sigma_{\mathrm{Ref}}/2$, where $\Delta_{\mathrm{Ref}} = m_1^{1/2} - m_2^{1/2}$ is the difference and $\Sigma_{\mathrm{Ref}} = m_1^{1/2} + m_2^{1/2}$ is the sum of the square roots of the masses of the two reference ions. $C_{\mathrm{TOF}} = (2t - t_1 - t_2)/[2(t_1 - t_2)]$ comprises all measured time-of-flight values t, t_1 and t_2 of the ion of interest and the reference ions, respectively. Thus, it relates the mass m of the ion of interest to the reference-ion masses $m_{1,2}$ and allows re-evaluation of the data if the value of the reference masses changes.

Our application of the MR-TOF MS method is the first for rare isotope beams. Figure 2a shows a typical time-of-flight spectrum of the mass-53 ions, which resulted from the addition of 47,000 single-shot spectra (experimental cycles) taken over a period of about 3.5 h. The typical timescale for an 'experimental cycle', that is, the time from proton impact, after which we collect an ion ensemble, until its ejection from the MR-TOF device and detection, is of the order of 10 ms. Figure 2b shows similar spectra for mass 54 in the form of a two-dimensional colour-coded intensity plot as a function of time-of-flight (abscissa) and spectrum number (ordinate), where each number corresponds to the accumulated data of a spectrum like that in Fig. 2a. For this particular series of measurements, the proton bombardment on the ISOLDE target was interrupted for spectra 12 to 15 to exclude the possibility that counts detected in the time region of ^{54}Ca$^+$ originated from any long-lived species. As expected, all short-lived species disappear from the spectrum (the production of stable ^{54}Cr$^+$ decreases as well). In addition, the highly selective laser ionization was switched off

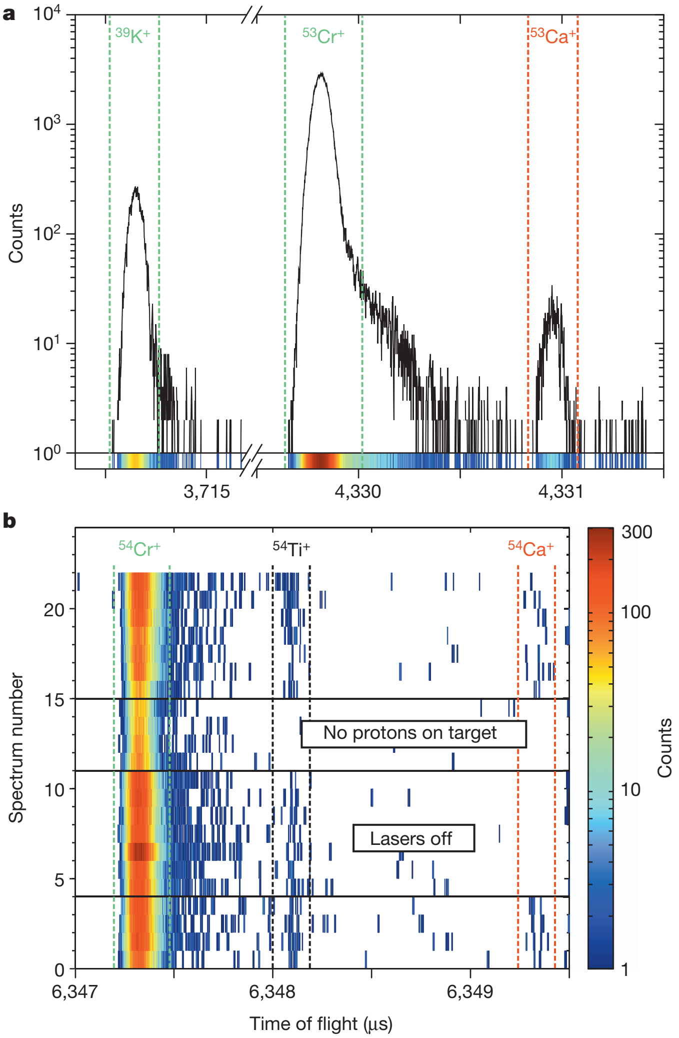


Figure 2 | Time-of-flight spectra. **a**, Time-of-flight spectrum of $A = 53$ nuclides delivered from ISOLDE (^{53}Cr$^+$, ^{53}Ca$^+$) and the reference ion ^{39}K$^+$ from the offline ion source. At bottom is the same spectrum compressed to a plot with colour-coded ion counts. **b**, Two-dimensional colour-coded intensity plot of time-of-flight spectra of $A = 54$ nuclides. The number of ion counts (colour coded, key at right) is shown as a function of time of flight on the abscissa and as a function of the measurement time (spectrum number) on the ordinate. Intensity plots are shown for different experimental conditions (with laser ionization on and protons on target, unless indicated otherwise).

Table 1 | Results of the calcium mass measurements

Isotope	$T_{1/2}$	Meas. type	Ref. nuclide(s)	r_{ICR}	C_{TOF}	Mass excess (keV/c^2)	
						ISOLTRAP	TITAN
^{51}Ca	10.0(8) s	ICR	^{39}K	1.3079136760(144)	NA	−36332.07(0.58)	−36338.9(22.7)
^{52}Ca	4.6(3) s	ICR	^{39}K	1.3336358720(184)	NA	−34266.02(0.71)	−34244.6(61.0)
		MR-TOF	^{39}K, ^{52}Cr	NA	0.501632110(785)	−34271.7(10.2)	
^{53}Ca	461(90) ms	MR-TOF	^{39}K, ^{53}Cr	NA	0.50184761(309)	−29387.8(43.3)	—
^{54}Ca	90(6) ms	MR-TOF	^{39}K, ^{54}Cr	NA	0.50210648(323)	−25161.0(48.6)	—

$T_{1/2}$, half-life[30]; measurement (meas.) type (ICR, ion cyclotron resonance; MR-TOF, multi-reflection time-of-flight mass spectrometry); reference (ref.) nuclide(s) used for the calibration; r_{ICR}, experimental frequency ratio; C_{TOF}, TOF constant; mass excess, $M_{\mathrm{exc}} = (M - Au)$, where M is the atomic mass, A is the atomic number and u is the unified atomic mass unit. For comparison, the TITAN[4] values are also listed. The mass values of the reference nuclides are $m(^{39}\mathrm{K}) = 38963706.4864(49)$ μu, $m(^{52}\mathrm{Cr}) = 51940506.26(63)$ μu, $m(^{53}\mathrm{Cr}) = 52940648.17(62)$ μu, $m(^{54}\mathrm{Cr}) = 53938879.18(61)$ μu (ref. 28). NA, not applicable.

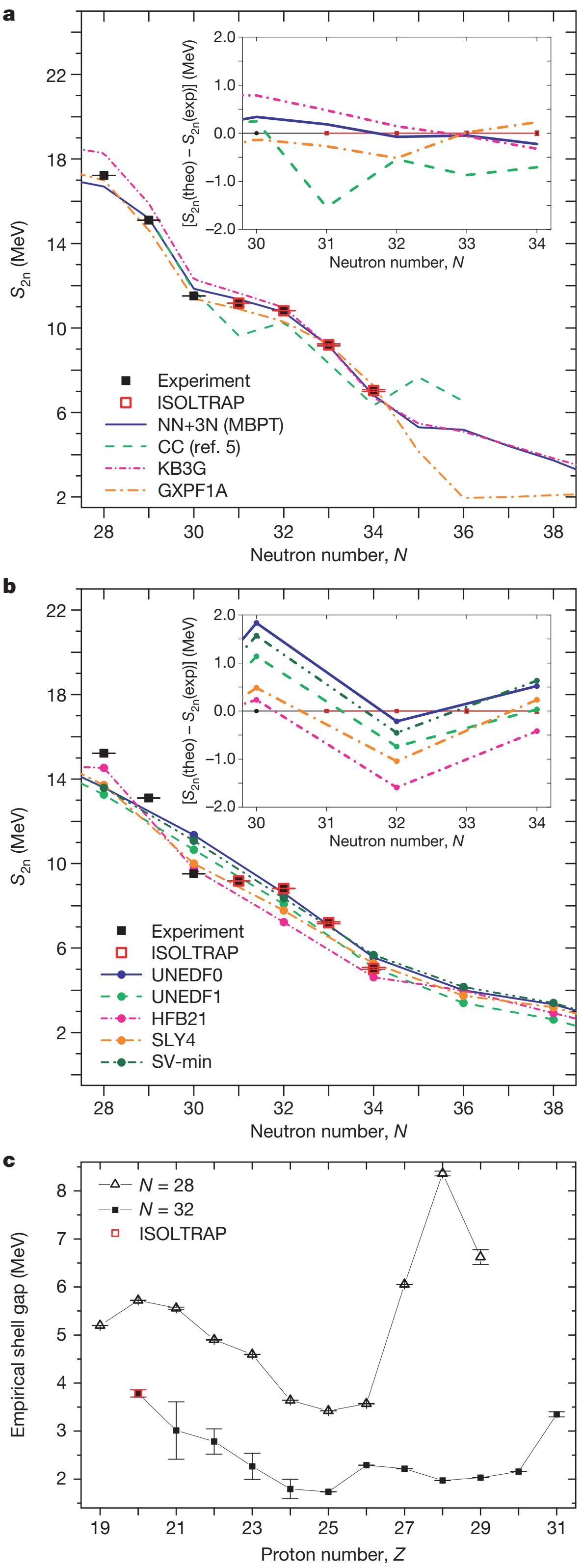


Figure 3 | Comparison of experimental results with theoretical predictions. **a**, **b**, Two-neutron separation energy S_{2n} (ref. 28) of the neutron-rich calcium isotopes as a function of neutron number N, where the new ISOLTRAP values are shown in red. In **a**, the ISOLTRAP masses are compared to predictions from microscopic valence-shell calculations with three-nucleon forces (NN+3N) based on chiral effective field theory (solid line, MBPT) and large-space coupled-cluster calculations including three-nucleon forces as density-dependent two-body interactions (dashed line, CC)[5]. For comparison, we also show the results of the phenomenological shell-model interactions KB3G[21] and GXPF1A[22]. In **b**, the ISOLTRAP masses are compared to state-of-the-art nuclear density-functional-theory predictions[15,29]. Insets in **a** and **b** show the difference between the theoretical predictions and experiment. **c**, Empirical two-neutron shell gap as a function of proton number Z for $N = 28$ and $N = 32$.

during spectra 5 to 11, which resulted in the disappearance of the ion counts in question. This unambiguously identified these ions as ^{54}Ca. Figure 2b corresponds to about 90 min of data-taking. MR-TOF MS spectra of ^{53}Ca and ^{54}Ca were taken in total for 12.6 h and 18.2 h, respectively.

Our results (r_{ICR} and C_{TOF}) for the exotic calcium isotopes investigated (^{51,52}Ca and ^{53,54}Ca, respectively) are summarized in Table 1, including the resulting mass excesses. The ISOLTRAP values of ^{51}Ca and ^{52}Ca determined with the Penning trap agree well with the recent measurements by TITAN[4]. The uncertainties were reduced by factors of 40 and 80, respectively, owing to longer excitation times (600 ms in the case of ISOLTRAP as compared to 80 ms in the case of TITAN), higher cyclotron frequencies and higher calcium ion yields. The masses of ^{53,54}Ca determined by the MR-TOF MS have been experimentally addressed for the first time. As a consistency check, the ^{52}Ca mass was also measured by the new MR-TOF method, and the mass excess is in full agreement with both Penning-trap results (Table 1). Furthermore, a second cross-check measurement in the vicinity of the newly measured masses was performed. The mass excess of the stable isotope ^{58}Fe was determined with the stable reference isotopes ^{58}Ni and ^{85}Rb. The measurement resulted in a mass excess of −62,168.0(47.0) keV/c^2, where the statistical uncertainty is given in parentheses. With a deviation of 13.5 keV/c^2 from the literature value[28], it agrees well within its statistical uncertainty. The uncertainties in the MR-TOF method quoted in Table 1 for ^{53}Ca and ^{54}Ca denote the statistical standard deviation. For the cross-checks, the MR-TOF method has thus been employed to measure the mass of a slightly lighter isotope and a slightly heavier isotope, ^{52}Ca and ^{58}Fe, respectively. The deviations from the Penning-trap measurement and the literature value, respectively, are taken as estimates of the relative systematic uncertainty, which lies in the low 10^{-7} range. Additional cross-check measurements to determine the systematic uncertainty have been performed over a wide mass range and will be detailed elsewhere. The precision and fast measurement cycle of the MR-TOF method makes this a promising approach for the mass spectrometry of isotopes with lower yield and shorter half-life than currently accessible.

The binding energies encode information about the ordering of shell occupation, and thus are essential in the quest for shell closures in exotic regions of the nuclear chart. Our high-precision data can be used to provide a critical benchmark for the behaviour far from stability, namely, the two-neutron separation energy $S_{2n} = B(Z,N) - B(Z,N-2)$, where $B(Z,N)$ is the binding energy (defined as positive) of a nucleus with Z protons and N neutrons. The S_{2n} values are a preferred probe of the evolution of nuclear structure with neutron number, and can be used to challenge model predictions, as shown in Fig. 3. The pronounced decrease in S_{2n} revealed by the new ^{53}Ca and ^{54}Ca ISOLTRAP masses is similar to the decrease beyond the doubly magic ^{48}Ca. In general, correlations induced by deformation could also cause such a reduction in S_{2n}, but in the calcium isotopes studied here deformation is expected to have no role[29]. Therefore, our new data unambiguously establish a prominent shell closure at $N = 32$. The strength of this shell closure can be evaluated from the two-neutron shell gap, that is, the two-neutron separation energy difference $S_{2n}(Z,N) - S_{2n}(Z,N+2)$. Figure 3c shows a two-neutron shell gap for ^{52}Ca of almost 4 MeV, where the rise towards ^{52}Ca at $N = 32$ is as steep as that towards ^{48}Ca at $N = 28$. The peaks at $N = Z$ in Fig. 3c are due to the additional correlation energy for symmetric $N = Z$ nuclei, known as Wigner energy.

Calcium marks the heaviest chain of isotopes studied with three-nucleon forces based on chiral effective field theory[3–6]. Figure 3a shows the predictions of our microscopic calculations with three-nucleon forces (that is, 'NN + 3N') using many-body perturbation theory

(MBPT) for the valence-neutron interactions[3,4]; it also shows predictions of large-scale coupled-cluster ('CC') calculations including the continuum and three-nucleon forces as density-dependent two-body interactions[5]. Confronted with the S_{2n} values obtained from the new masses of ^{53}Ca and ^{54}Ca, we find an excellent agreement with the predictions. We have also calculated perturbatively the effect of residual three-valence-neutron forces. This provides only a very small repulsive contribution, lowering S_{2n} by about 50 keV to 250 keV from ^{51}Ca to ^{54}Ca in the MBPT framework. The agreement of the NN+3N calculations with the new ISOLTRAP masses is remarkable, because their parameters are fitted only to the properties of few-nucleon systems while their level of accuracy here is similar to the phenomenological shell-model interactions KB3G[21] and GXPF1A[22], which are adjusted to the medium-mass region. The CC calculations predict a $N = 32$ shell gap very close to our measurements; however, the oscillations in S_{2n} for odd neutron numbers on either side of the shell closure disagree with experiment.

In Fig. 3b we compare the new ISOLTRAP masses to state-of-the-art nuclear density-functional-theory (DFT) predictions[15,29], which have recently provided global predictions for nuclear driplines[15]. (The dripline marks the limit of existence where nuclei cease to be bound.) This shows that modern DFT calculations can reproduce the masses of $^{52-54}$Ca. In particular, the UNEDF0 and SV-min functionals are in very good agreement. However, the DFT calculations predict an almost linear progression of the two-neutron separation energy that does not lead to the experimentally observed decrease in S_{2n} at ^{48}Ca and at ^{52}Ca, established by the new masses of ^{53}Ca and ^{54}Ca. In the NN+3N calculations, the neutron dripline is obtained at ^{62}Ca, but owing to the very flat behaviour of the binding energies past ^{60}Ca (refs 3, 5, 15, 29), the limit of calcium isotopes is very difficult to predict theoretically. In addition, the effects of the continuum, not included in the MBPT calculations, will be decisive for nuclides close to the dripline[5,29]. Note that the calcium isotopes studied here are still well bound, with 4.2 MeV separation energy in ^{54}Ca, so that the effects of the continuum on the ground-state energies are small (of the order of 100 keV)[5], comparable to the small effects of residual three-nucleon forces.

The present results based on precision mass measurements with a multi-reflection time-of-flight method reinforce the suggestion that pronounced structural effects are important in exotic nuclei and that shell effects do not smear out far from stability. Our results provide useful information for all theoretical models, and they show that a description of extreme neutron-rich nuclei can be closely connected to a deeper understanding of nuclear forces. Chiral effective field theory provides this connection and an exciting framework for exploring neutron-rich nuclei. The measurements of the ^{53}Ca and ^{54}Ca isotopes, accessed in this work, present anchor points to pin down nuclear forces. Finally, we note that the advantages of the MR-TOF method as compared to Penning-trap mass spectrometry will also be important for new experimental facilities, which will provide even more exotic ion beams. The present and future developments of low-energy beams at facilities for the study of exotic nuclides such as ARIEL, CARIBU, FAIR, FRIB, HIE-ISOLDE, RIBF and SPIRAL 2 will considerably extend the available range of rare isotopes towards the nuclear driplines. The minute production rates of isotopes with half-lives in the millisecond range and substantial isobaric contamination pose experimental challenges that are barely met by Penning traps now, but can be overcome with the MR-TOF method.

Received 19 December 2012; accepted 23 April 2013.

1. Baumann, T., Spyrou, A. & Thoennessen, M. Nuclear structure experiments along the neutron drip line. *Rep. Prog. Phys.* **75,** 036301 (2012).
2. Hammer, H.-W., Nogga, A. & Schwenk, A. Three-body forces: from cold atoms to nuclei. *Rev. Mod. Phys.* **85,** 197–217 (2013).
3. Holt, J. D. *et al.* Three-body forces and shell structure in calcium isotopes. *J. Phys. G* **39,** 085111 (2012).
4. Gallant, A. T. *et al.* New precision mass measurements of neutron-rich calcium and potassium isotopes and three-nucleon forces. *Phys. Rev. Lett.* **109,** 032506 (2012).
5. Hagen, G. *et al.* Evolution of shell structure in neutron-rich calcium isotopes. *Phys. Rev. Lett.* **109,** 032502 (2012).
6. Roth, R. *et al.* Medium-mass nuclei with normal-ordered chiral NN+3N interactions. *Phys. Rev. Lett.* **109,** 052501 (2012).
7. Wollnik, H. & Przewloka, M. Time-of-flight mass spectrometers with multiply reflected ion trajectories. *Int. J. Mass Spectrom. Ion Process.* **96,** 267–274 (1990).
8. Epelbaum, E., Hammer, H.-W. & Meißner, U.-G. Modern theory of nuclear forces. *Rev. Mod. Phys.* **81,** 1773–1825 (2009).
9. Warner, D. Not-so-magic numbers. *Nature* **430,** 517–519 (2004).
10. Sorlin, O. & Porquet, M.-G. Nuclear magic numbers: new features far from stability. *Prog. Part. Nucl. Phys.* **61,** 602–673 (2008).
11. Janssens, R. V. F. Unexpected doubly magic nucleus. *Nature* **459,** 1069–1070 (2009).
12. Otsuka, T. *et al.* Magic numbers in exotic nuclei and spin-isospin properties of the NN interaction. *Phys. Rev. Lett.* **87,** 082502 (2001).
13. Otsuka, T. *et al.* Three-body forces and the limit of oxygen isotopes. *Phys. Rev. Lett.* **105,** 032501 (2010).
14. Hagen, G. *et al.* Continuum effects and three-nucleon forces in neutron-rich oxygen isotopes. *Phys. Rev. Lett.* **108,** 242501 (2012).
15. Erler, J. *et al.* The limits of the nuclear landscape. *Nature* **486,** 509–512 (2012).
16. Janssens, R. V. F. *et al.* Structure of ^{52,54}Ti and shell closures in neutron-rich nuclei above ^{48}Ca. *Phys. Lett. B* 546, 55–62 (2002).
17. Mantica, P. F. *et al.* β decay of neutron-rich $^{53-56}$Ca. *Phys. Rev. C* **77,** 014313 (2008).
18. Crawford, H. L. *et al.* β decay and isomeric properties of neutron-rich Ca and Sc isotopes. *Phys. Rev. C* **82,** 014311 (2010).
19. Huck, A. *et al.* Beta decay of the new isotopes ^{52}K, ^{52}Ca, and ^{52}Sc; a test of the shell model far from stability. *Phys. Rev. C* **31,** 2226–2237 (1985).
20. Gade, A. *et al.* Cross-shell excitation in two-proton knockout: structure of ^{52}Ca. *Phys. Rev. C* **74,** 021302(R) (2006).
21. Poves, A. *et al.* Shell model study of the isobaric chains A=50, A=51 and A=52. *Nucl. Phys. A* **694,** 157–198 (2001).
22. Honma, M. *et al.* New effective interaction for pf-shell nuclei and its implications for the stability of the N=Z=28 closed core. *Phys. Rev. C* **69,** 034335 (2004).
23. Blaum, K. High-accuracy mass spectrometry with stored ions. *Phys. Rep.* **425,** 1–78 (2006).
24. Schweikhard, L. & Bollen, G. (eds) Special issue on ultra-accurate mass spectrometry and related topics. *Int. J. Mass Spectrom.* **251,** 85–312 (2006).
25. Mukherjee, M. *et al.* ISOLTRAP: an on-line Penning trap for mass spectrometry on short-lived nuclides. *Eur. Phys. J. A* **35,** 1–29 (2008).
26. Wolf, R. N. *et al.* On-line separation of short-lived nuclei by a multi-reflection time-of-flight device. *Nucl. Instrum. Methods A* **686,** 82–90 (2012).
27. Fedosseev, V. N. *et al.* Upgrade of the resonance ionization laser ion source at ISOLDE on-line isotope separation facility: new lasers and new ion beams. *Rev. Sci. Instrum.* **83,** 02A903 (2012).
28. Wang, M. *et al.* The AME2012 atomic mass evaluation. *Chinese Phys. C* **36,** 1603–2014 (2012).
29. Forssén, C. *et al.* Living on the edge of stability, the limits of the nuclear landscape. *Phys. Scr. T* **152,** 014022 (2013).
30. Audi, G. *et al.* The NUBASE2012 evaluation of nuclear properties. *Chinese Phys. C* **36,** 1157–1286 (2012).

Acknowledgements This work was supported by the BMBF (contracts 06GF9102, 05P12HGCI1, 05P12HGFNE, 06DA70471, 06DD9054), the DFG (grants SFB 634 and GE2183/2-1), the ERC (grant 307986 STRONGINT), the EU through ENSAR (grant 262010), the Helmholtz Alliance HA216/EMMI, the French IN2P3, the ISOLDE Collaboration and the Max-Planck Society. Computations were performed at the Jülich Supercomputing Center.

Author Contributions D.B., Ch.B., R.B.C., S.K., D.L., V.M., D.N., M.R., J. Stanja, F.W. and R.N.W. performed the experiment. V.M. and F.W. performed the data analysis. J.D.H., J.M., A.S. and J. Simonis performed the NN+3N (MBPT) calculations. K.B., S.K., D.L., A.S., L.S. and F.W. prepared the manuscript. All authors discussed the results and contributed to the manuscript at all stages.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.W. (wienholtz@physik.uni-greifswald.de).

LETTER

doi:10.1038/nature12190

Defect pair separation as the controlling step in homogeneous ice melting

Kenji Mochizuki[1], Masakazu Matsumoto[2] & Iwao Ohmine[3]

On being heated, ice melts into liquid water. Although in practice this process tends to be heterogeneous, it can occur homogeneously inside bulk ice[1]. The thermally induced homogeneous melting of solids is fairly well understood, and involves the formation and growth of melting nuclei[1–5]. But in the case of water, resilient hydrogen bonds render ice melting more complex. We know that the first defects appearing during homogeneous ice melting are pairs of five- and seven-membered rings, which appear and disappear repeatedly and randomly in space and time in the crystalline ice structure[6–8]. However, the accumulation of these defects to form an aggregate is nearly additive in energy, and results in a steep free energy increase that suppresses further growth. Here we report molecular dynamics simulations of homogeneous ice melting that identify as a crucial first step not the formation but rather the spatial separation of a defect pair. We find that once it is separated, the defect pair—either an interstitial (I) and a vacancy (V) defect pair (a Frenkel pair), or an L and a D defect pair (a Bjerrum pair)[9]—is entropically stabilized, or 'entangled'. In this state, defects with threefold hydrogen-bond coordination persist and grow, and thereby prepare the system for subsequent rapid melting.

Ice melting trajectory calculations were performed by modelling the superheating of crystalline ice Ih (ref. 10) and following the melting process (Methods). Figure 1a plots the potential energy per molecule along a typical trajectory. Figure 1b shows the corresponding change in the total number of 'off-lattice' water molecules (n), and in the size of the largest cluster of such molecules (n_{LC}) that grows into the melting nucleus. (For snapshots of the corresponding hydrogen-bond structures, see Supplementary Fig. 1.) We define water molecules as 'off-lattice' if they are more than 0.1 nm away from the nearest lattice point of the ice structure (Supplementary Fig. 2), and consider off-lattice molecules within a distance of 0.6 nm as adjacent to each other and forming a cluster, and thereby determine n_{LC}.

In the quiescent period ($<$2,150 ps, Fig. 1), we see in the ice hydrogen-bond network structural defects of five- and seven-membered rings ('5+7 defect') and/or pairs of L and D defects ('L+D complex') randomly scattered in space[6–8]. (In the ordered ice structure, there is one proton between two oxygen atoms; in the L and D defects, there are respectively no protons and two protons between two oxygen atoms.) Although the appearance of these 5+7 defects and/or L+D complexes is the first step in ice melting, their simple accumulation will not result in melting: because they retain fourfold hydrogen-bond coordination, individual water molecules in the 5+7 defects are strongly restricted in their motion, and entropy will therefore not increase rapidly with the energy increase. Hence, the free energy will rise sharply with an increase in the number of these defects. In the trajectories involving only 5+7 defects and/or L+D complexes, we indeed found that n hardly exceeds 15, and that the system repeatedly exhibits intermittent creation and annihilation of small-sized melting clusters.

In trajectories resulting in melting, the defect growing to form the melting nucleus is either an I defect spatially separated from its accompanying V defect, or a D defect spatially separated from its accompanying L defect (see Supplementary Fig. 3 for defect structures). In these separated structures, the I defect encompasses an additional lattice water molecule, while the D defect has two hydrogen atoms between two oxygen atoms and thus breaks the Bernal-Fowler ice rule[11]. Separated defects form occasionally during the recrystallization that occurs after thermal fluctuations have created several 5+7 defects and/or L+D complexes. Although defects usually appear and disappear rapidly, recrystallization occasionally fails to revert back to ice obeying the Bernal-Fowler rule and instead yields separated I and V (or D and L) defects while the rest of the region recovers the original crystalline structure. Clusters of 5+7 defects (or L+D complexes) with $n_{\mathrm{LC}} > 5$ accumulate in our simulations every 275 ps on average, with 14% forming a separated defect pair and the rest recrystallizing. Once a

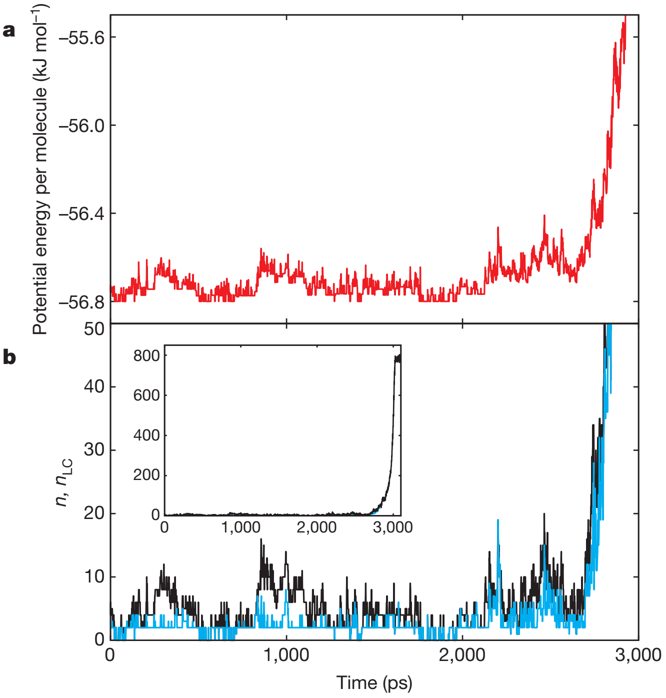


Figure 1 | Potential energy per molecule and number of off-lattice molecules in a melting trajectory. **a**, Potential energy per molecule of the inherent structures of a typical melting trajectory. At 0 ps, a temperature jump from 270 K to 275 K is imposed. **b**, Main panel, total number n of off-lattice water molecules (black line) and size n_{LC} of the largest cluster (blue line) for the trajectory of **a**. Note that the melting point of the water model (TIP4P) is $T_{\mathrm{m}} = 232$ K, and 275 K is a superheated state. The potential energy fluctuates during the long quiescent period owing to intermittent creation and annihilation of small melting clusters. This lasts until 2,150 ps, when relatively large energy fluctuations appear with the creation of a separated I–V pair. Rapid growth of the melting cluster, and thus rapid energy rise, starts only at 2,730 ps. Inset, data plotted using n up to 800, that is, almost complete melting.

[1] School of Physical Sciences, The Graduate University for Advanced Studies (SOKENDAI), Myodaiji, Okazaki 444-8585, Japan. [2] Graduate School of Natural Science and Technology, Okayama University, 3-1-1 Tsushima, Okayama 700-8530, Japan. [3] Institute for Molecular Science, Myodaiji, Okazaki 444-8585, Japan.

separated defect pair is created (for example, at 2,150 ps in Fig. 1), it undergoes facile dislocation on the lattice (see below), and the distance between I and V (or D and L) defects increases rapidly. Such a separated pair is hard to annihilate because its recombination would require the right sequence of many hydrogen-bond alternations (Supplementary Information section 3), and we thus refer to it as entangled. A typical separated I–V defect pair is shown Fig. 2a.

The small melting cluster containing an I defect just after separation from its V defect often exhibits rapid motion in the lattice, which for the trajectory shown in Fig. 1 lasts from 2,150 to 2,730 ps (see also Supplementary Videos 1 and 2 and Supplementary Fig. 4). After this induction period of about 600 ps, the system takes very little time to reach the liquid state (about 0.3 ns in Fig. 1). The average time from initial I–V separation to total melting is about 1 ns, with a Poisson-type time distribution peaking around $t = 0.5$ ns. The fast formation of a critical nucleus and subsequent melting is a direct result of separated I (or D) defects enabling facile hydrogen-bond alternations in water ice (see also below).

To quantify the degree of disorder of the hydrogen-bond network of the melting nucleus, we define the topological 'edit' distance d_T as the

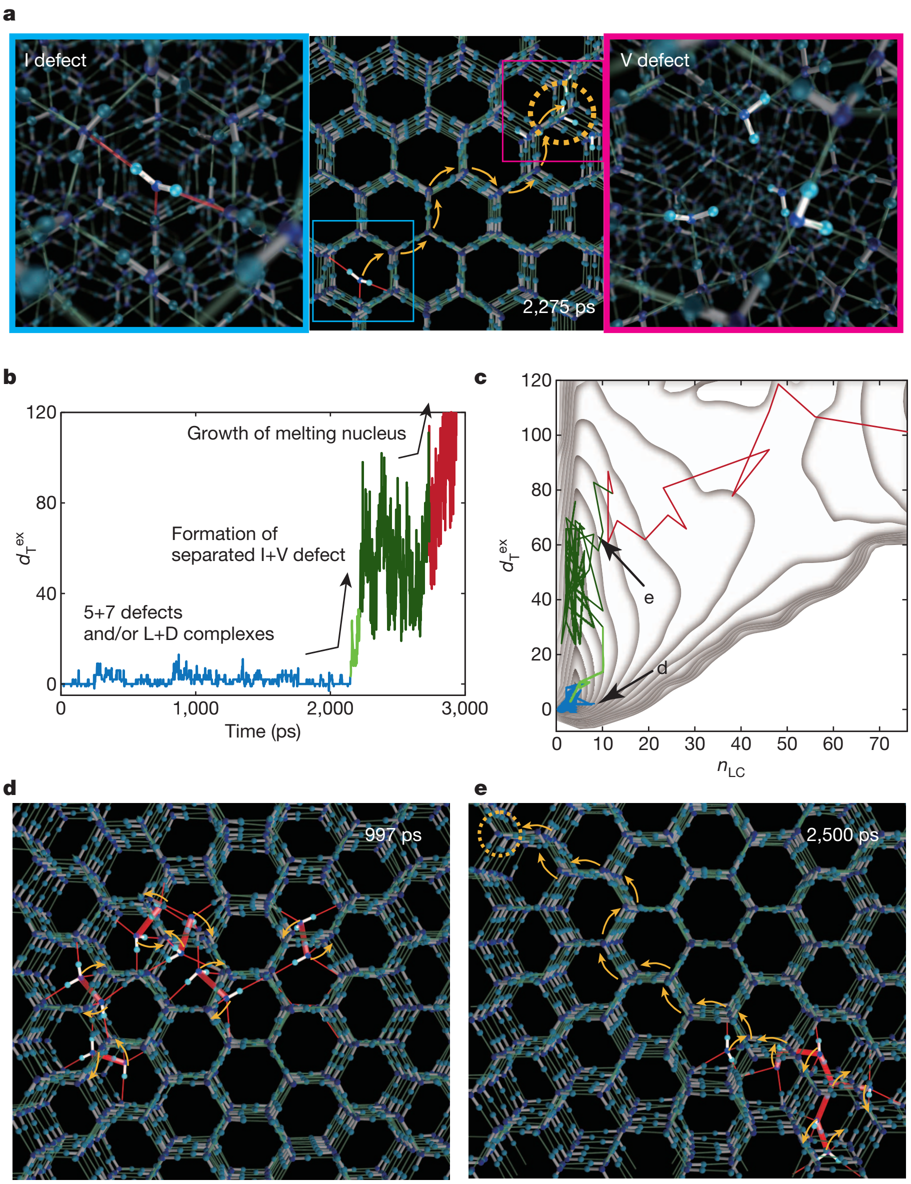


Figure 2 | Snapshots of hydrogen-bond structures and the time evolution of excess edit distance, d_T^{ex}. **a**, Typical example of a separated I–V defect pair in ice (middle panel). The I defect and the four molecules surrounding the V defect (located at the centre of the dotted circle) are indicated by bright colours and shown magnified in the left and right panels, respectively. The 'editing' path to recover a crystalline ice structure (see main text) is indicated by yellow arrows. For this example, with only a pair of I–V defects, d_T^{ex} is 58. **b**, Evolution of d_T^{ex}, calculated every 1 ps, for the trajectory of Fig. 1. The quiescent period containing 5+7 defects and/or L+D complexes, the formation of the separated I–V defect and the subsequent induction period, and the final period where the melting cluster reaches the size of the critical nucleus and grows rapidly, are coloured blue, green and red, respectively. These structurally distinct periods are further emphasized in **c**, which shows d_T^{ex} against the largest cluster size n_{LC} (calculated every 10 ps). Separation of the defect pair is signalled by the rapid increase in d_T^{ex}, while n_{LC} remains small throughout the subsequent induction period lasting from $t = 2{,}150$ to 2,730 ps (green lines). In both **b** and **c**, dark-green shading for $t = 2{,}210$–2,730 ps indicates when separated defects rapidly dislocate in ice. The thick grey contour lines in the background are the free-energy contours shown in Fig. 3a. **d**, **e**, Snapshots of hydrogen-bond structures from the same melting trajectory, taken at 997 ps and 2,500 ps (indicated by black arrows in **c**). Red lines indicate hydrogen bonds to off-lattice molecules. The yellow arrows indicate the edit paths that recover a crystalline structure through the formation, cutting and directional inversion of hydrogen bonds (see text and Supplementary Fig. 5). The structure in **d** contains mostly 5+7 defects, and that in **e** a separated I–V defect pair.

minimum number of all hydrogen-bond additions and deletions (that is, edits) needed to recover from a given disordered hydrogen-bond structure to the network topology of the closest proton-disordered ice structure[12] (Supplementary Information section 2.1 gives the detailed procedure for estimating d_T.) Formation of a 5+7 defect, for example, introduces two off-lattice molecules and increases d_T by 4 (sometimes 6) because it requires the breaking and also the creation of 2 hydrogen bonds to recover the original ice structure. When 5+7 defect pairs accumulate, d_T increases by $2n$ and we can then define the excess d_T (denoted as d_T^{ex}) as d_T minus $2n$. d_T^{ex} is a measure of the difficulty of resolving hydrogen-bond disorder, and hence of the degree of hydrogen-bond network entanglement.

Figure 2b shows the time evolution of d_T^{ex} for the melting trajectory in Fig. 1. d_T^{ex} stays small when melting clusters consisting only of 5+7 defects and/or L+D complexes appear intermittently in the quiescent period, and increases suddenly at 2,150 ps when a separated I–V pair is created. The melting nucleus containing the I defect then changes its position and size for about 600 ps, with d_T^{ex} fluctuating around larger values. The nucleus finally starts growing rapidly at 2,730 ps (Fig. 1), and d_T^{ex} increases again. In this growing process, an I defect often couples with its surrounding 5+7 defects to rapidly convert ice hydrogen-bond network structures into liquid structures. The plot of d_T^{ex} against n_{LC} in Fig. 2c shows that the melting trajectory is initially characterized by $d_T^{ex} < 10$ as n_{LC} fluctuates between 0 and 5 (n fluctuates between 0 and 15), and that d_T^{ex} then increases and fluctuates more strongly after a separated I–V pair is created. With the increase of d_T^{ex}, the defect pair becomes harder to remove via annihilation. The benefit of a quantitative measure of network disorder is also illustrated by the hydrogen-bond network structures arbitrarily selected from the trajectory of Fig. 1 and shown in Fig. 2d and e: although the structure in Fig. 2d may look more distorted than in Fig. 2e, the latter hydrogen-bond network has the larger d_T^{ex} value and is thus characterized by larger network disorder.

We calculate the free energy of the system from the melting trajectories, and plot it against n_{LC} and d_T^{ex} in Fig. 3a. (The procedure used to estimate the free energy is described in Supplementary Information section 2.2). The contours are such that they enforce a strong direction dependence on the minimum free-energy path to melting: d_T^{ex} first increases owing to the formation of a separated pair defect, and only then does n_{LC}, the size of the largest cluster that goes on to form the melting nucleus, increase. The solid-state free-energy minimum is found at $n_{LC} = 4$, which is about the average size of the melting cluster in the quiescent period. The critical nucleus, that is, the saddlepoint between solid and liquid, is located around $n_{LC} = 50$ and is about 18 kJ mol^{-1} higher in energy than the solid state. Beyond the critical nucleus size, the free energy decreases monotonically to the liquid state. The size of the critical nucleus decreases with increasing temperature, with our calculations predicting that it is around $n_{LC} = 33$ at 280 K.

This contour map can be projected on n_{LC} to obtain a one-dimensional free-energy surface. The resultant free-energy curve F_W (grey line in Fig. 3b) is then decomposed into the free energy of the 'un-separated

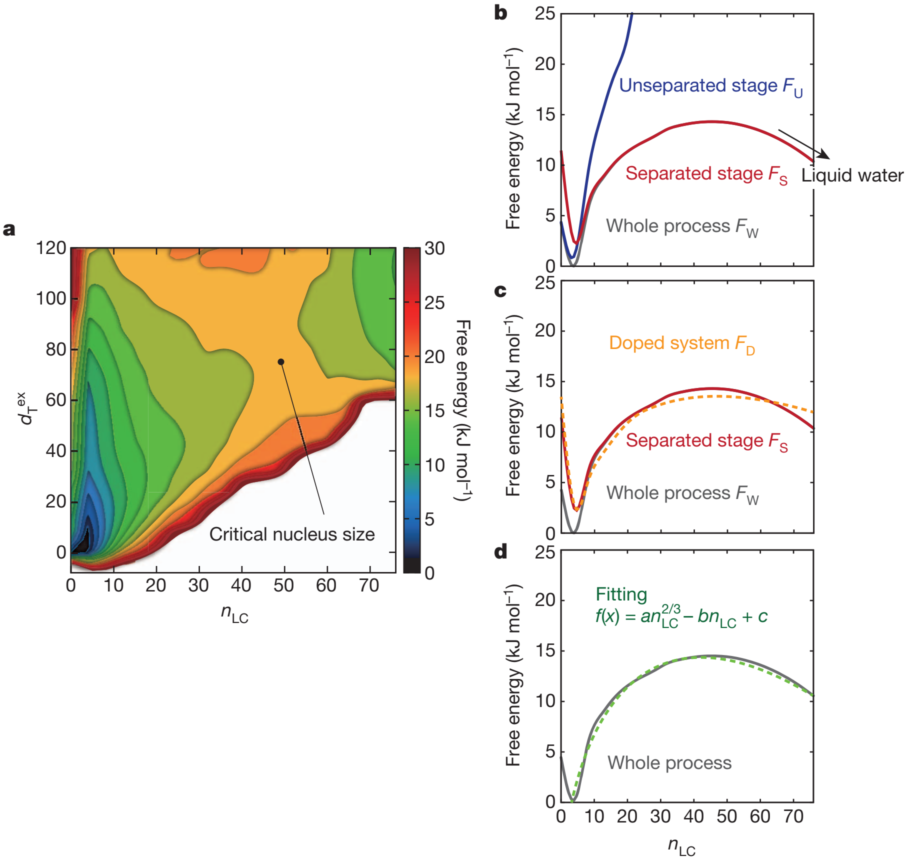


Figure 3 | Free energies. **a**, The contour map of free energy as a function of largest cluster size (n_{LC}) and excess edit distance (d_T^{ex}) at 275 K. The origin (n_{LC}, d_T^{ex}) = (0,0) represents ice with no defects. The critical nucleus at (n_{LC}, d_T^{ex}) = (49, 78) is indicated by a black dot. **b**, The contour map in **a** is projected onto n_{LC} to give the 'whole' process curve (F_W, grey line). The whole process is divided into the unseparated stage (F_U: blue line) and the separated stage (F_S: red line), see main text. **c**, Free energy of the doped system (F_D: orange dashed line), which mimics ice containing a permanent I defect (see main text). Note that F_D is almost identical to F_S (the free energy of the separated stage). **d**, The empirical function $f(n) = an^{2/3} - bn + c$ (green dashed line) of classical nucleation theory with the parameters $a = 5.93$, $b = 1.13$ and $c = -9.71$, optimized by approximating F_W. The classical free energy curve overlaps well with F_W.

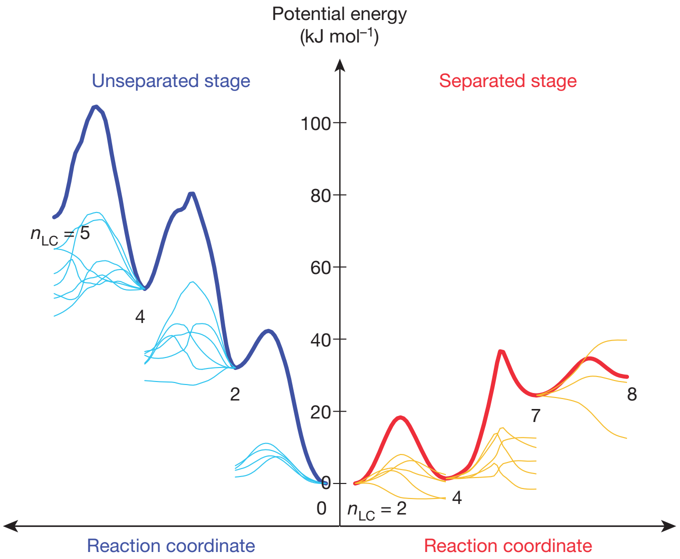


Figure 4 | Potential energy surfaces during defect growth. Illustration of how potential energy changes as defect sizes n_{LC} increase, in the unseparated (left: blue lines) and separated (right: red lines) stage. Total potential energy changes are plotted in bold lines, and potential energy changes of individual molecules directly involved in defect growth are plotted with thin blue (left) or thin yellow (right) lines. Potential energies are along reaction coordinates (see Methods). The nucleus size n_{LC} is indicated by a number at each minimum of an inherent structure. The starting points of individual molecular potential energy changes are shifted to the corresponding minima of the total energy surfaces. For example, five water molecules are directly involved in the defect growth from $n_{LC} = 2$ to 4 in the separated stage (five yellow curves at the bottom of the right side figure); the curve corresponding to the separated I defect exhibits a monotonic decay.

stage' (F_U) and that of the 'separated stage' (F_S), with $F_W = -k_B T \ln[\exp(-\beta F_U) + \exp(-\beta F_S)]$, where k_B is Boltzmann's constant, T absolute temperature and $\beta = 1/k_B T$. F_U is calculated by projecting the 'un-separated' part of the contour in Fig. 3a with $d_T^{ex} \leq 10$, and F_S by projecting the other part of the contour. Figure 3b shows that F_U increases monotonically and sharply with n_{LC}. In contrast, the initial slope of F_S is less than half that of F_U, showing the facile growth of the melting nucleus after defect separation. F_S intersects with F_U at around $n_{LC} = 6$.

We next performed separate molecular dynamics simulations for a system with an extra water molecule added to ice. The I defect of this so-called 'doped system' cannot be annihilated because there is no V-defect counterpart, so it mimics a long-lived I defect. Figure 3c shows that the free energy of the doped system (F_D) is almost identical to F_S, with this strong resemblance between the two free energies illustrating the key role of separated I-defect formation in melting.

The functional form of free energy in classical nucleation theory[13] is given by $f(n_{LC}) = an_{LC}^{2/3} - bn_{LC} + c$, where n_{LC} and $n_{LC}^{2/3}$ correspond to the volume and the surface area of the melting nucleus, respectively. Figure 3d illustrates that F_W can be fitted well with a function $f(n_{LC})$ (green line), indicating that ice melting can be described within the classical nucleation theory framework when looking at this stochastic part of the melting process. However, homogeneous ice melting advances under normal super-heated conditions only when a separated I–V (or separated L–D) pair is involved, and this early molecular melting process is much more intricate and elaborate than the simple stochastic process considered by classical nucleation theory. Although also stochastic in nature, it requires at least a two-dimensional description of the free energy, as in Fig. 3a. But once a defect pair has been separated and entanglement created, melting is seen to proceed in a near-classical stochastic manner and further destroys the hydrogen-bond network of the crystal.

Figure 4 plots the total potential energy changes[14] along a typical melting trajectory as either a separated I defect or un-separated defect clusters grow, along with the potential energy changes of individual water molecules involved in defect growth. Energy barriers[15] are particularly low when an I defect is involved in separated cluster growth, as its dangling bonds change the hydrogen-bond coordination with other water molecules in the lattice to reduce its own energy while destabilizing other water molecules. The total potential energy of the system thus slowly increases with the increase of n_{LC}. The average energy needed to create an additional defect, $\Delta U(n_{LC}) = U(n_{LC}+1) - U(n_{LC})$, is only about 4–5 kJ mol^{-1} per off-lattice molecule for $n_{LC} = 5$ to 20. For comparison, the experimentally obtained value for liquid water (that is, $n_{LC} = \infty$) at 0 °C is 6 kJ mol^{-1}. But in the absence of separated I (or D) defects, the energies of all individual molecules directly involved in defect formation and growth increase significantly with the breaking of their hydrogen bonds to form strongly hindered four hydrogen-bond coordinations. The total energy steadily increases with n_{LC}, and $\Delta U(n_{LC})$ ranges from 9 to 13 kJ mol^{-1} per off-lattice molecule for $n_{LC} = 5$ to 10. This value is much larger than that in the presence of the separated I defect.

The average entropy term $T\Delta S(n_{LC})$ for creating one additional defect in the lattice is about 4 kJ mol^{-1} for $n_{LC} = 5$ to 10, and then gradually increases with further growth of the liquid fragments[16] (Supplementary Fig. 6), whereas $\Delta U(n_{LC})$ remains nearly constant at about 5 kJ mol^{-1} for $n_{LC} > 5$. This results in the difference between ΔU and $T\Delta S$ gradually decreasing, and in $T\Delta S$ surpassing ΔU at the critical nucleus size, so the free-energy surface has a gentle and convex upward slope up to the critical nucleus size (Fig. 3b). This feature of the potential energy surface and the fact that it is smooth with small energy barriers ensure that melting proceeds rapidly once separated I defects have formed and the system has gone through the subsequent short induction period.

We note that although we have shown that homogeneous melting under normal super-heating conditions only proceeds when separated I–V (or separated D–L) defect pairs are created, very high temperatures will simply induce total collapse of the ice network[1]. But separated defect pairs may possibly also play a role during the very late stages of water freezing[17–20], as separated I–V defects can induce hydrogen-bond reorientations that stabilize the system as a more proton-ordered ice[6], after an overall crystalline structure has been attained.

METHODS SUMMARY

Molecular dynamics trajectory calculations were performed on a system containing 896 water molecules in an almost cubic cell (edge lengths are 3.142, 3.110 and 2.932 nm) with periodic boundary conditions. We use the TIP4P water model[21], which is one of the most successful models in terms of reproducing the thermodynamic properties of water[22]. Intermolecular interaction is smoothly truncated from 1.0530 nm to 1.1638 nm. A preparatory calculation of the proton-disordered ice Ih structure evolving for 1 ns at 250 K is followed by step-by-step increase of the temperature; a pre-melting equilibration run is performed for 1 ns at 270 K after a 1 ns run at 260 K, then the melting process is observed for several nanoseconds at 275 K. More than 10 μs of trajectories in total are used for statistical analyses. Even though 275 K is higher than the melting temperature $T_m = 232$ K (ref. 23) predicted by the TIP4P model, 275 K is found to be about the lowest temperature at which the system melts[24]. At a higher temperature, for example 300 K, the crystal collapses as soon as it is heated up, while at 275 K the system shows an induction time of a few nanoseconds before melting starts.

Hundreds of trajectories are started with different initial proton order configurations and different initial velocities applied to water molecular motions in ice. Temperature is controlled by the Nose-Hoover thermostat[25]. The density, instead of the pressure, of the system is kept constant in this work. The density is set to 0.935 g cm^{-3}, which is the density of ice in this model at the melting point under atmospheric pressure. We mainly focus on the molecular mechanism of the initial stage of melting, when the volume of the system has not yet changed substantially.

In order to find energy barriers required for the structural changes of the hydrogen-bond network, reaction coordinate analyses[15] are performed on inherent structures of the system. The inherent structures are obtained by applying the conjugate gradient method[26] to the instantaneous structures visited by the trajectories.

Received 30 April 2012; accepted 12 April 2013.

1. Iglev, H., Schmeisser, M., Simeonidis, K., Thaller, A. & Laubereau, A. Ultrafast superheating and melting of bulk ice. *Nature* **439,** 183–186 (2006).
2. Fecht, H. J. Defect-induced melting and solid-state amorphization. *Nature* **356,** 133–135 (1992).
3. Cahn, R. W. Materials science: Melting from within. *Nature* **413,** 582–583 (2001).
4. Forsblom, M. & Grimvall, G. How superheated crystals melt. *Nature Mater.* **4,** 388–390 (2005).
5. Jin, Z. H., Gumbsch, P., Lu, K. & Ma, E. Melting mechanisms at the limit of superheating. *Phys. Rev. Lett.* **87,** 055703 (2001).
6. Tanaka, H. & Mohanty, J. On the Debye-Waller factor of hexagonal ice: a computer simulation study. *J. Am. Chem. Soc.* **124,** 8085–8089 (2002).
7. Grishina, N. & Buch, V. Structure and dynamics of orientational defects in ice I. *J. Chem. Phys.* **120,** 5217–5225 (2004).
8. Donadio, D., Raiteri, P. & Parrinello, M. Topological defects and bulk melting of hexagonal ice. *J. Phys. Chem. B* **109,** 5421–5424 (2005).
9. Bjerrum, N. Structure and properties of ice. *Science* **115,** 385–390 (1952).
10. McBride, C., Vega, C., Sanz, E., MacDowell, L. G. & Abascal, J. L. F. The range of meta stability of ice-water melting for two simple models of water. *Mol. Phys.* **103,** 1–5 (2005).
11. Bernal, J. D. & Fowler, R. H. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *J. Chem. Phys.* **1,** 515–548 (1933).
12. Petrenko, V. F. & Whitworth, R. W. *Physics of Ice* (Oxford Univ. Press, 1999).
13. Abraham, F. F. *Homogeneous Nucleation Theory: The Pretransition Theory of Vapor Condensation* (Academic, 1974).
14. Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge Univ. Press, 2004).
15. Henkelman, G., Uberuaga, B. P. & Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113,** 9901–9904 (2000).
16. Matsumoto, M., Baba, A. & Ohmine, I. Topological building blocks of hydrogen bond network in water. *J. Chem. Phys.* **127,** 134504 (2007).
17. Matsumoto, M., Saito, S. & Ohmine, I. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. *Nature* **416,** 409–413 (2002).
18. Jacobson, L. C., Hujo, W. & Molinero, V. Amorphous precursors in the nucleation of clathrate hydrates. *J. Am. Chem. Soc.* **132,** 11806–11811 (2010).
19. Walsh, M. R., Koh, C. A., Sloan, E. D., Sum, A. K. & Wu, D. T. Microsecond simulations of spontaneous methane hydrate nucleation and growth. *Science* **326,** 1095–1098 (2009).
20. Moore, E. B. & Molinero, V. Structural transformation in supercooled water controls the crystallization rate of ice. *Nature* **479,** 506–508 (2011).
21. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).
22. Jorgensen, W. L. & Madura, J. D. Temperature and size dependence for Monte-Carlo simulations of TIP4P water. *Mol. Phys.* **56,** 1381–1392 (1985).
23. Vega, C., Sanz, E. & Abascal, J. L. F. The melting temperature of the most common models of water. *J. Chem. Phys.* **122,** 114507 (2005).
24. Jacobson, L. C., Hujo, W. & Molinero, V. Thermodynamic stability and growth of guest-free clathrate hydrates: a low-density crystal phase of water. *J. Phys. Chem. B* **113,** 10298–10307 (2009).
25. Nose, S. Constant-temperature molecular dynamics. *J. Phys. Condens. Matter* **2,** SA115–SA119 (1990).
26. Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, 2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. J. Wales, N. Kosugi, H. Tanaka and S. Saito for their support and advice, and T. Yagasaki for performing the free-energy calculation to determine the melting temperature. Some of the calculations were carried out at the Research Center for Computational Science of NINS, Okazaki, Japan.

Author Contributions K.M. performed the simulations and carried out data analysis. M.M. developed the methods and carried out data analysis. All authors designed the study, discussed the results and contributed to writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.M. (kmochi@ims.ac.jp).

LETTER

doi:10.1038/nature12278

The importance of feldspar for ice nucleation by mineral dust in mixed-phase clouds

James D. Atkinson[1], Benjamin J. Murray[1], Matthew T. Woodhouse[2], Thomas F. Whale[1], Kelly J. Baustian[1], Kenneth S. Carslaw[1], Steven Dobbie[1], Daniel O'Sullivan[1] & Tamsin L. Malkin[1]

The amount of ice present in mixed-phase clouds, which contain both supercooled liquid water droplets and ice particles, affects cloud extent, lifetime, particle size and radiative properties[1,2]. The freezing of cloud droplets can be catalysed by the presence of aerosol particles known as ice nuclei[2]. One of the most important ice nuclei is thought to be mineral dust aerosol from arid regions[2,3]. It is generally assumed that clay minerals, which contribute approximately two-thirds of the dust mass, dominate ice nucleation by mineral dust, and many experimental studies have therefore focused on these materials[1,2,4–6]. Here we use an established droplet-freezing technique[4,7] to show that feldspar minerals dominate ice nucleation by mineral dusts under mixed-phase cloud conditions, despite feldspar being a minor component of dust emitted from arid regions. We also find that clay minerals are relatively unimportant ice nuclei. Our results from a global aerosol model study suggest that feldspar ice nuclei are globally distributed and that feldspar particles may account for a large proportion of the ice nuclei in Earth's atmosphere that contribute to freezing at temperatures below about −15 °C.

Pure cloud droplets remain liquid until cooled to the homogeneous freezing threshold at around 237 K (ref. 2). At warmer temperatures, freezing can be catalysed by heterogeneous ice nuclei. Heterogeneous nucleation can occur in several ways depending upon temperature and humidity[1,2]. Field observations and modelling studies of mixed-phase cloud formation have shown that supercooled liquid water is usually required before significant in-cloud ice formation occurs[2,8]. Therefore, in this study we focus on heterogeneous ice nucleation by mineral dust particles immersed in supercooled water droplets.

Atmospheric mineral dusts are inorganic particles of rock and soil that have been lifted into the atmosphere, predominantly from arid regions such as the Sahara[9]. Mineral dusts from these regions are considered an important source of ice nuclei in mixed-phase clouds, owing to their nucleation efficiency[7,10,11] and abundance in the atmosphere[9]. The importance of mineral dusts as ice nuclei is also supported by their number concentrations within atmospheric ice crystals, which are enhanced relative to the background aerosol[12]. At present the components in mineral dust responsible for ice nucleation are very poorly constrained. Although atmospheric dust concentrations and mineralogy vary spatially and temporally[9,13] (Supplementary Fig. 1), a large fraction of observed atmospheric dust mass around the world is made up of just a few minerals. Individual minerals are classified by their crystal structure and chemistry and can be identified with diffraction techniques[2]. The mineralogical composition of dust sampled from the atmosphere is shown in Supplementary Table 1. The clay minerals contribute approximately two-thirds of dust mass (kaolinites, 13%; montmorillonites, 2%; chlorites, 3%; micaceous minerals, such as the illites, 44%), with quartz (16%), feldspars (sodium/calcium feldspars (Na/Ca-feldspar), 8%; potassium feldspar (K-feldspar), 3%) and calcite (3%) responsible for much of the remainder.

Previous studies have investigated the ice-nucleating behaviour of dusts sampled from arid source regions or dusts selected as proxies for natural dust[1,2]. Studies of ice nucleation by individual minerals of varying purity immersed in water have focused on the clay minerals[1,2,4–6]. However, minerals are rarely available in a pure state, and quantification of secondary minerals associated with a particular sample is often neglected. Such characterization is necessary because a minor component may dominate ice nucleation. In this study, we present measurements of ice nucleation by samples of individual minerals in which the impurities were quantified using X-ray diffraction.

To determine the ice-nucleating behaviour of each mineral, we used an established droplet freezing technique[4,7]. Hundreds of micrometre-sized droplets containing a known amount of solid material were cooled at 1 K min^{-1}, and freezing was monitored by optical microscopy. The freezing temperatures of individual 14–16-μm-diameter droplets containing a range of different minerals are shown in Fig. 1a. In experiments with similar dust surface areas, the temperature at which

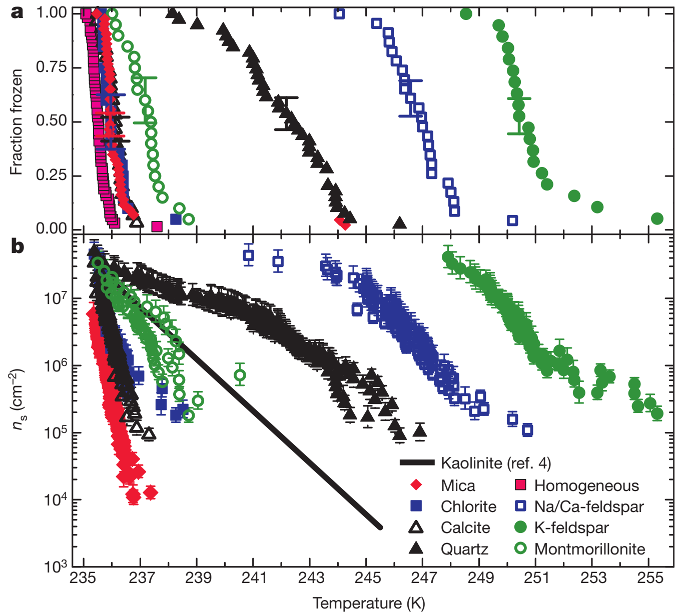


Figure 1 | Experimental freezing results for the individual minerals. **a**, Fraction of droplets, 14–16 μm in diameter and containing a range of mineral dusts, frozen as a function of temperature during cooling. An experiment in which droplets contained no solid inclusion and froze homogeneously is shown for comparison. Temperature uncertainty (not shown) is estimated at ±0.2 K. The indicated uncertainty in the measured fraction frozen is the root mean squared error (68% confidence limit) determined from the pairwise differences between the data and the fraction frozen calculated from the log-linear best fits to the n_s values. **b**, Nucleation site densities for droplets between 9 and 19 μm in diameter collected into four 2.5-μm-wide bins. The uncertainty in n_s is primarily due to droplet size measurements, and temperature uncertainty is as in **a**. The kaolinite parameterization is from ref. 4.

[1]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK. [2]Commonwealth Scientific and Industrial Research Organisation (CSIRO) Marine and Atmospheric Research, PMB 1, Aspendale, Victoria 3195, Australia.

50% of droplets were frozen was 250.5 K for K-feldspar, 247 K for Na/Ca-feldspar, 242.5 K for quartz and less than 237.5 K for the clay minerals and calcite. These results suggest that it is the minerals of the feldspar group, in particular K-feldspar, that make mineral dust particles effective immersion-mode ice nuclei in the atmosphere. This contrasts with the prevailing view[1,2], which is that clay minerals are the most important component of atmospheric mineral dust for ice nucleation.

Droplet freezing temperatures are dependent on experimental parameters such as droplet volume and mineral surface area, and are therefore of limited value[2]. To normalize the efficiency with which a material nucleates ice, we determine the number of nucleation sites per unit surface area[2,11,14], n_s (Fig. 1b and Supplementary Information). This method of quantifying ice-nucleation efficiency neglects the role of time dependence in nucleation on the basis that the particle-to-particle variability of ice nuclei is more important than the time dependence of nucleation[2,11,14,15]. Our derived n_s values for 9–19-µm-diameter droplets are shown in Fig. 1b. These data show that the feldspar minerals, in particular K-feldspar, are the most efficient mineral dust ice nuclei per unit surface area.

In airborne dusts, the abundance of clay minerals tends to be greater than the abundance of the feldspars, and it is therefore not immediately clear which minerals dominate ice nucleation in the atmosphere. The n_s values presented in Fig. 1b were combined with the average mineralogical composition of atmospheric dust to estimate the temperature-dependent ice nuclei concentration (Fig. 2). We have assumed that all particles are spherical to estimate their surface area and have made two limiting calculations, one assuming that dust particles are internally mixed (that is, each particle contains all eight minerals) and the other assuming they are externally mixed (each particle is composed of an individual mineral). The mixing state of atmospheric dust is poorly constrained but falls between these two limiting cases[16]. Despite accounting for only 3% of atmospheric dust by mass, K-feldspar dominates the number of ice nuclei above 248 K in both the internally mixed and externally mixed cases. One potential caveat to this conclusion is that clay mineral particles may have a smaller particle size than feldspar or quartz[13], and therefore may have a greater surface area per unit mass, which would increase the concentration of clay ice nuclei. However, even if the surface area of the clays were 100 times higher (probably an overestimate[7]), the feldspars would remain the dominant ice-nucleating minerals (Supplementary Fig. 4).

Because clouds glaciate over a wider range of temperatures than is achievable in the experiments presented above, it is important to quantify the nucleating efficiency of atmospheric ice nuclei over a broader temperature range. For example, an ice nuclei concentration of as few as 10^{-5} cm^{-3} in a cloud at ~266 K may be able to trigger substantial glaciation through the Hallett–Mossop ice multiplication process[17], and so measurements are needed at these temperatures. To extend the data for K-feldspar to this temperature regime, we performed a series of experiments with larger droplets, allowing for much larger particle surface areas per droplet and correspondingly lower values of n_s to be quantified. Results for K-feldspar using droplets 1 µl in volume are shown in Fig. 3. These data extend the range of experimental data up to 268 K. Combining the two techniques allowed us to determine n_s values over a range of eight orders of magnitude. The data show that the ice-nucleation efficiency of feldspar is measurable at temperatures relevant for the Hallett–Mossop process (about 265–270 K).

In Fig. 3, we compare our K-feldspar results with cloud-chamber-derived n_s values for a range of natural mineral dusts sampled from arid source regions[10]. Because feldspar is a major component of Earth's crust[2], it is ubiquitous in soils around the globe. K-feldspar makes up as much as ~24% by mass of soils throughout the Asian and African dust belt[18] and is also present in airborne dust in concentrations ranging from a few per cent[13] to 25% (ref. 19). We have estimated n_s values for natural dust in Fig. 3 by assuming that between 1% and 25% of the surface area of the dusts tested in ref. 10 was feldspar. In the temperature range where the data sets overlap, the agreement is very good.

Results presented here may explain discrepancies in existing experimental data for ice nucleation by mineral dusts. For example, a kaolinite sample from the Clay Mineral Society[4] (CMS) had much lower n_s values

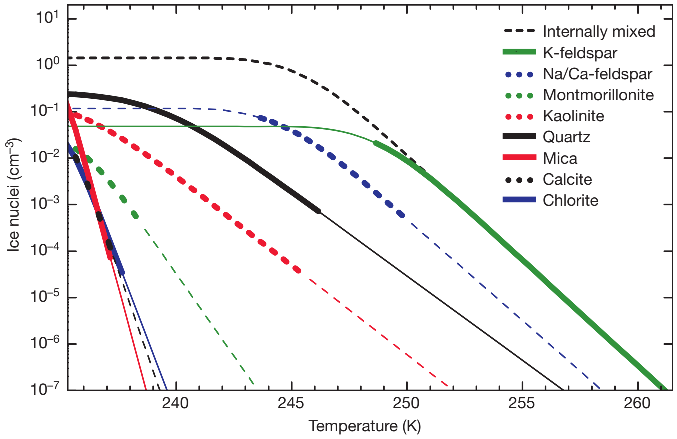


Figure 2 | Concentration of ice nuclei due to various minerals, for externally and internally mixed cases. Ice nuclei concentrations were estimated using the abundance of various minerals from Supplementary Table 1, taking the annually and globally averaged dust concentration (1.4 cm^{-3}) and size distribution from the GLOMAP model together with the n_s values in Fig. 1b (the K-feldspar line uses the n_s parameterization in Fig. 3). Lines with specific mineral names refer to individual minerals in an externally mixed case. Thick lines denote the range of experimental data and thin lines denote extrapolations outside this range.

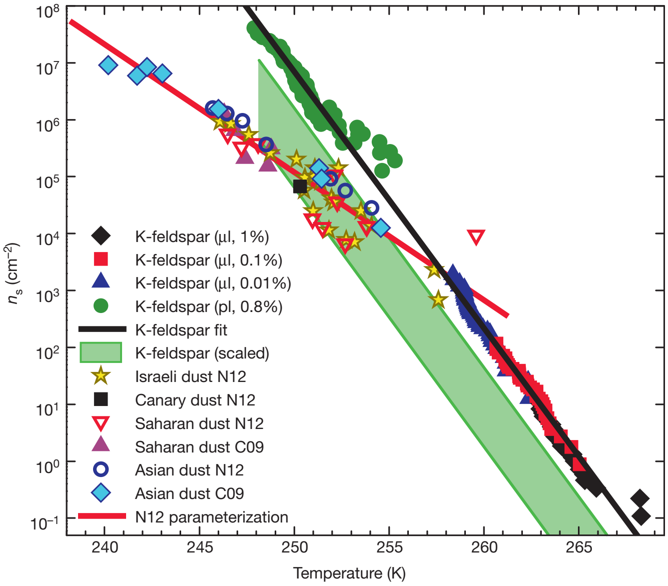


Figure 3 | Nucleation site density for K-feldspar and natural dusts. Data from Fig. 1b (picolitre experiments) extended to higher temperatures by use of microlitre-sized droplets, with a fit provided ($\ln(n_s) = -1.038T + 275.26$, valid between 248 and 268 K). Experimental K-feldspar concentrations in weight percent are provided in the key. Temperature uncertainties for microlitre experiments (not shown) were estimated at ±0.4 K and uncertainty in n_s (not shown) is estimated at ±25%. We also compare with data for several natural dust samples from ref. 10 (N12) and ref. 11 (C09). The mineralogical compositions of the dusts used in those references are unknown, but feldspar mass content in natural soils typically varies between 1% and ~25% (see text). Hence, we have scaled our n_s values assuming K-feldspar is present at between 1% and 25% of the natural dust particles' surface.

than a sample obtained from Sigma-Aldrich[2,5]. In this study, we have determined the mineralogical composition of these dusts and shown that the Sigma-Aldrich kaolinite contained 4.5% K-feldspar whereas the CMS kaolinite contained no detectable feldspar (Supplementary Table 2). It is likely that many individual particles of Sigma-Aldrich kaolinite contained feldspar, as is the case in atmospheric dust particles[16]. We have also quantified the mineralogy of the four montmorillonite samples used in a previous study of ice nucleation[6] (see Supplementary Table 2 for sample mineralogy). Three of the samples (M SWy-2, M KSF and M K-10) contain K-feldspar and had higher freezing temperatures than the M STx-1b sample, which did not contain measurable quantities of feldspar. Arizona test dust had the highest freezing temperature of any mixed-mineral dust proxy[6] and also contained the most K-feldspar (~20 wt%). In general, the more feldspar a sample contains, the higher the freezing temperature. We propose that the feldspar component controlled the nucleation of ice in these experiments, highlighting the need to characterize sample mineralogy in such work.

The mineralogical composition of soils in arid regions around the world varies substantially. Therefore, to quantify the global contribution of feldspars to ice nuclei concentrations it is necessary to use a global aerosol model. Global maps of dust number concentration, feldspar mass fraction and ice nuclei concentration from an aerosol and chemical transport model, GLOMAP, are shown in Fig. 4a–d. To calculate ice nuclei concentration, the mineral-resolved size distribution of dust was combined with our parameterization for the ice-nucleation efficiency of feldspar at 253 K (ice nuclei concentrations at other temperatures are shown in Supplementary Fig. 7). In addition, we also make the assumption that the minerals in mineral dust aerosol are externally mixed; this assumption produces a better match to the observational ice nuclei data at lower temperatures than does the opposing internally mixed assumption (Supplementary Fig. 8), although in reality the mixing state of atmospheric dust will lie somewhere between the internally mixed and externally mixed states.

The model data show larger ice nuclei concentrations close to the major dust sources in North Africa and Asia, which results in Northern Hemisphere dust ice nuclei concentrations around one to two orders of magnitude larger than in the Southern Hemisphere (comparing similar latitudes; Fig. 4c, d). Stratiform clouds in the Southern Hemisphere typically glaciate at lower temperatures[20–22], consistent with a lower concentration of ice nuclei than in the Northern Hemisphere.

To investigate the importance of dust mineralogy for modelling ice nuclei concentrations, we compared ice nuclei concentration in Fig. 4c with that predicted by a parameterization for natural dusts sampled from arid source regions[10]. Figure 4e shows that the two parameterizations are in agreement close to dust sources. However, the natural dust parameterization predicts ice nuclei concentrations up to 70% higher in regions remote from sources. This higher prediction arises because feldspar is more common in the larger-particle-size fractions (>2 μm) and therefore sediments out more rapidly than the minerals in the small-size fractions[13] (Fig. 4b). Hence, atmospheric mineral dust becomes less efficient at nucleating ice during transport through a non-chemical ageing process. Our results may also help to explain the chemical ageing process of dust ice nuclei, which is known to reduce the ice-nucleating efficiency of dust[23]. Feldspars are susceptible to emissions of acid gases such as SO_2, which can convert the surface of feldspar grains to clay minerals[24]. This may block ice-nucleation sites and reduce the efficiency of feldspar as an ice nucleus, which provides an explanation for the observed sensitivity of mineral dust ice nuclei to acid processing[23].

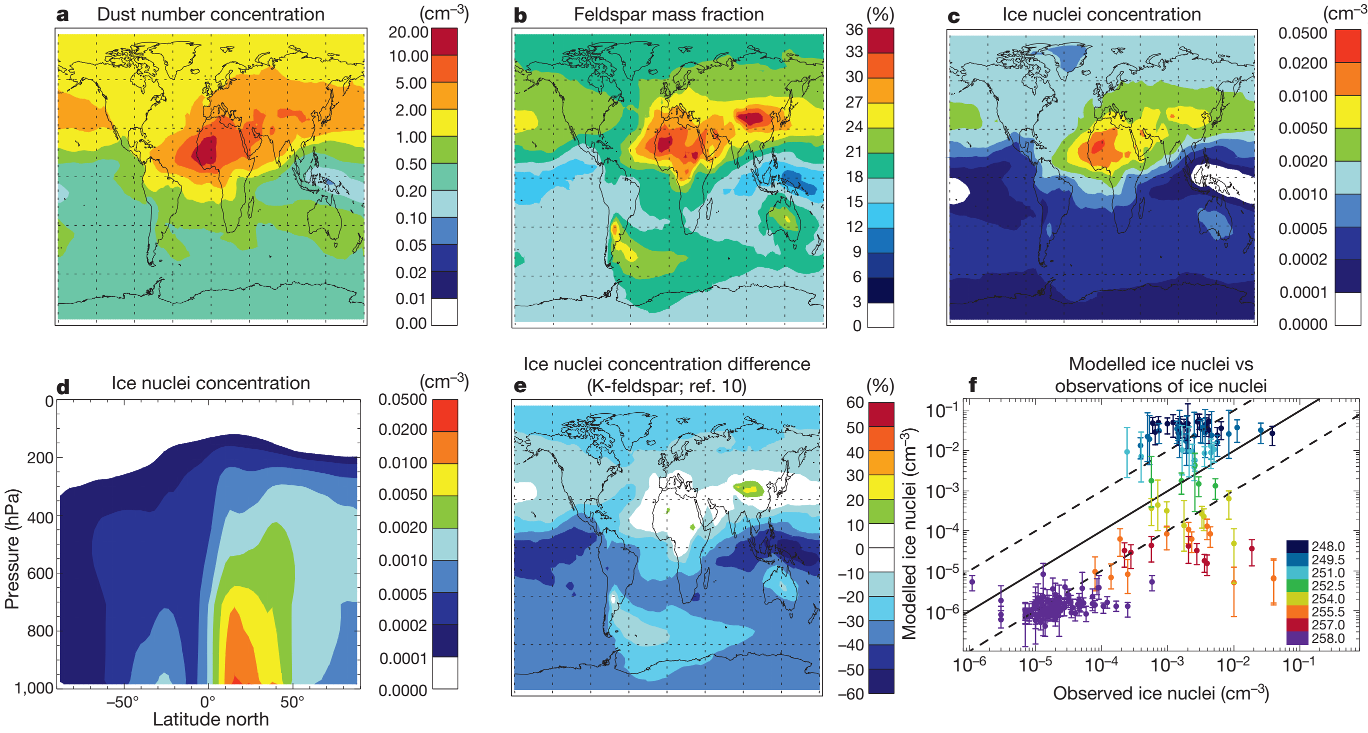


Figure 4 | Dust aerosol modelling study results. **a**, Modelled dust number concentrations. **b**, Total feldspar mass fraction of dust. **c**, Ice nuclei concentration due to K-feldspar at 253 K, calculated using our n_s parameterization and modelled particle surface areas. On the basis of observations in Supplementary Table 1, we assume 35% of feldspar mass is K-feldspar. **d**, Latitudinal zonal mean values of ice nuclei from **c**. **e**, Comparison of **c** versus ice nuclei concentrations calculated using a mineralogy-independent parameterization based on desert dust samples at 253 K (ref. 10). **f**, Comparison of model ice nuclei concentrations from K-feldspar mineral dust with field measurements of total ice nuclei. Annual mean modelled ice nuclei concentrations are taken at the same pressure level as the field observation, with the observation temperature used to calculate n_s. Only observations between 248 and 258.15 K are shown; a comparison at higher temperatures is shown in Supplementary Fig. 9. Vertical error bars represent the maximum and minimum modelled monthly mean values. See Supplementary Table 4 for field campaign details. Parts **a**–**c** and **e** use concentrations at an altitude where the pressure is 600 hPa.

We also compare GLOMAP mineral dust ice nuclei concentrations with field measurements of ice nuclei concentrations (where the aerosol-processing temperature was ≤258 K) from around the world in Fig. 4f. The data are scattered around the 1:1 line, indicating that feldspar is one of the most important ice nuclei in Earth's atmosphere. The model tends to over-predict ice nuclei concentrations at temperatures below ~249 K. It is important to note that we do not include nucleation scavenging, where dust is removed when it serves as ice nuclei or cloud condensation nuclei, and we may therefore over-predict dust concentrations in regions remote from the source. Also, the cluster of data at 258 K, from a ship-borne study around southern Australasia (Supplementary Table 4), is consistently below the 1:1 line. This may indicate that in addition to mineral dust, other ice nuclei sources were also important in this region. At temperatures higher than 258 K, feldspar mineral dust is much less important as ice nuclei and cannot account for the observed ice nuclei concentrations (Supplementary Fig. 9). At these warmer temperatures, other types of ice nuclei, possibly of biogenic origin[25], may become increasingly important.

Finally, recent work suggests that human activity has led to a substantial increase in atmospheric dust concentrations and that the sources of this dust have changed[9,26]. Because potential dust sources around the world have very different feldspar contents[18], changes in the location of dust sources may have consequences for the concentration of ice nuclei in the atmosphere and the associated aerosol indirect effect.

METHODS SUMMARY

Experimental method. Picolitre experiments (Fig. 1) were performed using a freezing assay of micrometre-sized droplets in a manner similar to previously described[4,7]. The main variation between this and the previous method is the development of a new cold stage configuration to improve the thermal stability and control of the system, with improvements to temperature measurements reducing uncertainty to ±0.2 K. The microlitre experiments were performed with a separate system in which cooling, temperature measurement and control were provided by a Stirling engine-powered flat-plate chiller (Grant-Asymptote EF600). Droplets were then deposited onto a large, siliconized glass slide using a pipette, and freezing was monitored using a digital camera without magnification. The analysis of the data from the picolitre and microlitre experiments was identical. Mineral dust characterization methods have been described previously[7].

Modelling with GLOMAP. GLOMAP is a size- and composition-resolving two-moment microphysical aerosol scheme[27], run within the TOMCAT chemical transport model[28]. GLOMAP has previously been used to study atmospheric processing of mineral dust[29]. The model is driven by reanalysis meteorology for the year 2000. GLOMAP was extended to represent eight mineral types, as specified by the surface mineralogy map of ref. 18. Dust is represented in 12 size bins, ranging in diameter from 0.1 to >20.0 μm. Dust emissions are prescribed from AEROCOM recommendations for the year 2000[30]. The experimentally derived n_s values were combined with modelled atmospheric particle sizes, composition and concentrations (annual global averages at the altitude corresponding to a pressure of 600 hPa) to estimate ice nuclei concentrations (Fig. 2). Two particle mixing states were considered. In the internally mixed case, all particles contain the considered minerals in the same ratios, whereas in the externally mixed case individual particles are composed of single minerals with the overall population composition controlled by the dust mineralogy.

Received 30 January; accepted 7 May 2013.
Published online 12 June 2013.

1. Hoose, C. & Möhler, O. Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments. *Atmos. Chem. Phys.* **12,** 9817–9854 (2012).
2. Murray, B. J., O'Sullivan, D., Atkinson, J. D. & Webb, M. E. Ice nucleation by particles immersed in supercooled cloud droplets. *Chem. Soc. Rev.* **41,** 6519–6554 (2012).
3. DeMott, P. J. *et al.* African dust aerosols as atmospheric ice nuclei. *Geophys. Res. Lett.* **30,** 1732 (2003).
4. Murray, B. J., Broadley, S. L., Wilson, T. W., Atkinson, J. D. & Wills, R. H. Heterogeneous freezing of water droplets containing kaolinite particles. *Atmos. Chem. Phys.* **11,** 4191–4207 (2011).
5. Lüönd, F., Stetzer, O., Welti, A. & Lohmann, U. Experimental study on the ice nucleation ability of size-selected kaolinite particles in the immersion mode. *J. Geophys. Res.* **115,** D14201 (2010).
6. Pinti, V., Marcolli, C., Zobrist, B., Hoyle, C. R. & Peter, T. Ice nucleation efficiency of clay minerals in the immersion mode. *Atmos. Chem. Phys.* **12,** 5859–5878 (2012).
7. Broadley, S. L. *et al.* Immersion mode heterogeneous ice nucleation by an illite rich powder representative of atmospheric mineral dust. *Atmos. Chem. Phys.* **12,** 287–307 (2012).
8. de Boer, G., Morrison, H., Shupe, M. D. & Hildner, R. Evidence of liquid dependent ice nucleation in high-latitude stratiform clouds from surface remote sensors. *Geophys. Res. Lett.* **38,** L01803 (2011).
9. Ginoux, P., Prospero, J. M., Gill, T. E., Hsu, N. C. & Zhao, M. Global-scale attribution of anthropogenic and natural dust sources and their emission rates based on MODIS Deep Blue aerosol products. *Rev. Geophys.* **50,** RG3005 (2012).
10. Niemand, M. *et al.* A particle-surface-area-based parameterization of immersion freezing on desert dust particles. *J. Atmos. Sci.* **69,** 3077–3092 (2012).
11. Connolly, P. J. *et al.* Studies of heterogeneous freezing by three different desert dust samples. *Atmos. Chem. Phys.* **9,** 2805–2824 (2009).
12. Pratt, K. A. *et al.* In situ detection of biological particles in cloud ice-crystals. *Nat. Geosci.* **2,** 398–401 (2009).
13. Glaccum, R. A. & Prospero, J. M. Saharan aerosols over the tropical north-Atlantic: mineralogy. *Mar. Geol.* **37,** 295–321 (1980).
14. Vali, G. Quantitative evaluation of experimental results an the heterogeneous freezing nucleation of supercooled liquids. *J. Atmos. Sci.* **28,** 402–409 (1971).
15. DeMott, P. J. Quantitative descriptions of ice formation mechanisms of silver iodide-type aerosols. *Atmos. Res.* **38,** 63–99 (1995).
16. Jeong, G. Y. Bulk and single-particle mineralogy of Asian dust and a comparison with its source soils. *J. Geophys. Res.* **113,** D02208 (2008).
17. Crawford, I. *et al.* Ice formation and development in aged, wintertime cumulus over the UK: observations and modelling. *Atmos. Chem. Phys.* **12,** 4963–4985 (2012).
18. Nickovic, S., Vukovic, A., Vujadinovic, M., Djurdjevic, V. & Pejanovic, G. High-resolution mineralogical database of dust-productive soils for atmospheric dust modeling. *Atmos. Chem. Phys.* **12,** 845–855 (2012).
19. Kandler, K. *et al.* Ground-based off-line aerosol measurements at Praia, Cape Verde, during the Saharan Mineral Dust Experiment: microphysical properties and mineralogy. *Tellus* **63B,** 459–474 (2011).
20. Choi, Y.-S., Lindzen, R. S., Ho, C.-H. & Kim, J. Space observations of cold-cloud phase change. *Proc. Natl Acad. Sci. USA* **107,** 11211–11216 (2010).
21. Kanitz, T. *et al.* Contrasting the impact of aerosols at northern and southern midlatitudes on heterogeneous ice formation. *Geophys. Res. Lett.* **38,** L17802 (2011).
22. Hoose, C., Kristjánsson, J. E., Chen, J.-P. & Hazra, A. A classical-theory-based parameterization of heterogeneous ice nucleation by mineral dust, soot, and biological particles in a global climate model. *J. Atmos. Sci.* **67,** 2483–2503 (2010).
23. Sullivan, R. C. *et al.* Irreversible loss of ice nucleation active sites in mineral dust particles caused by sulphuric acid condensation. *Atmos. Chem. Phys.* **10,** 11471–11487 (2010).
24. Zhu, C., Veblen, D. R., Blum, A. E. & Chipera, S. J. Naturally weathered feldspar surfaces in the Navajo Sandstone aquifer, Black Mesa, Arizona: Electron microscopic characterization. *Geochim. Cosmochim. Acta* **70,** 4600–4616 (2006).
25. Burrows, S. M., Hoose, C., Pöschl, U. & Lawrence, M. G. Ice nuclei in marine air: biogenic particles or dust? *Atmos. Chem. Phys.* **13,** 245–267 (2013).
26. Mahowald, N. M. *et al.* Observed 20th century desert dust variability: impact on climate and biogeochemistry. *Atmos. Chem. Phys.* **10,** 10875–10893 (2010).
27. Spracklen, D. V., Pringle, K. J., Carslaw, K. S., Chipperfield, M. P. & Mann, G. W. A global off-line model of size-resolved aerosol microphysics: I. Model development and prediction of aerosol properties. *Atmos. Chem. Phys.* **5,** 2227–2252 (2005).
28. Arnold, S. R., Chipperfield, M. P. & Blitz, M. A. A three-dimensional model study of the effect of new temperature-dependent quantum yields for acetone photolysis. *J. Geophys. Res.* **110,** D22305 (2005).
29. Shi, Z. B. *et al.* Minor effect of physical size sorting on iron solubility of transported mineral dust. *Atmos. Chem. Phys.* **11,** 8459–8469 (2011).
30. Dentener, F. *et al.* Emissions of primary aerosol and precursor gases in the years 2000 and 1750 prescribed data-sets for AeroCom. *Atmos. Chem. Phys.* **6,** 4321–4344 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Cuadros, L. Neve and R. Finch for help sourcing mineral samples, P. DeMott for providing ice nuclei observational data, and T. W. Wilson for discussions. We acknowledge the European Research Council (FP7, 240449 ICE) and the Natural Environment Research Council (NE/I013466/1, NE/I020059/1, NE/I019057/1) for funding. K.S.C. is a Royal Society Wolfson Research Merit Award holder.

Author Contributions J.D.A. conducted the picolitre experiments, analysed the data and wrote the paper, and T.F.W. performed and analysed the microlitre experiments. K.J.B. and D.O. contributed to the experimental study, and S.D. helped draft the manuscript. M.T.W. led the global modelling study in collaboration with K.S.C. T.L.M. did the X-ray analysis of the mineral samples. B.J.M. oversaw the project and helped to write the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.J.M. (b.j.murray@leeds.ac.uk).

doi:10.1038/nature12168

A Jurassic avialan dinosaur from China resolves the early phylogenetic history of birds

Pascal Godefroit[1], Andrea Cau[2], Hu Dong-Yu[3,4], François Escuillié[5], Wu Wenhao[6] & Gareth Dyke[7]

The recent discovery of small paravian theropod dinosaurs with well-preserved feathers in the Middle–Late Jurassic Tiaojishan Formation of Liaoning Province (northeastern China)[1–4] has challenged the pivotal position of *Archaeopteryx*[3,4], regarded from its discovery to be the most basal bird. Removing *Archaeopteryx* from the base of Avialae to nest within Deinonychosauria implies that typical bird flight, powered by the forelimbs only, either evolved at least twice, or was subsequently lost or modified in some deinonychosaurians[3,5]. Here we describe the complete skeleton of a new paravian from the Tiaojishan Formation of Liaoning Province, China. Including this new taxon in a comprehensive phylogenetic analysis for basal Paraves does the following: (1) it recovers it as the basal-most avialan; (2) it confirms the avialan status of *Archaeopteryx*; (3) it places Troodontidae as the sister-group to Avialae; (4) it supports a single origin of powered flight within Paraves; and (5) it implies that the early diversification of Paraves and Avialae took place in the Middle–Late Jurassic period.

Theropoda Marsh, 1881
Maniraptora Gauthier, 1986
Paraves Sereno, 1997
Avialae Gauthier, 1986
Aurornis xui gen. et sp. nov.

Etymology. *Aurora*, Latin for daybreak, dawn; *Ornis*, Greek for bird; *xui*, in honour of Xu Xing, for his exceptional and continuing contribution to our understanding of the evolution and biology of feathered dinosaurs.

Holotype. Yizhou Fossil and Geology Park (YFGP)-T5198, a complete articulated skeleton with associated integumentary structures.

Locality and horizon. Yaolugou, Jianchang, western Liaoning Province, China; Middle–Late Jurassic Tiaojishan Formation (see Supplementary Information).

Diagnosis. Manual phalanx I-1 distinctly more robust than the radius; robust postacetabular process of ilium not markedly deflected ventrally and with a horizontal dorsal margin; distal end of ischium dorsoventrally expanded and formed by a hook-like ventral process delimiting a prominent distal obturator notch and by a longer dorsal distal process; metatarsal I gracile and elongate (about 30% of metatarsal III length) (see Supplementary Information for differential diagnosis).

Description. The holotype and only currently known specimen of *A. xui* (YFGP-T5198) is 51 cm in length (Fig. 1 and Supplementary Fig. 4) and was probably an adult individual; its frontals are fused, neurocentral sutures of all visible vertebrae are closed, and the astragalus–calcaneum complex is completely fused to the tibia. As in *Anchiornis* and *Eosinopteryx*, the skull is slightly shorter than the femur. The snout of *Aurornis* is about half the basal length of the skull, proportionally longer than in *Eosinopteryx* and *Mei*[6] (Supplementary Table 2), and lower in lateral view than in *Xiaotingia*[3] (Fig. 2a, b). In contrast to *Anchiornis* and *Mei*, the nares of *Aurornis* do not extend beyond the rostral border of the antorbital fenestra[2]. The maxillary process of the premaxilla is long, slender and contacts the nasal, excluding the maxilla from the external naris; in *Archaeopteryx* and *Anchiornis*, the maxillary process of the premaxilla is short and the maxilla participates in the ventral margin of the external naris[3]. The maxillary fenestra is large, separated from the antorbital fenestra by a narrow interfenestral bar, as in *Anchiornis*[2]. The premaxillary fenestra is larger than in *Anchiornis* and has a ventral margin located below that of the maxillary fenestra. The jugal is more gracile than in *Anchiornis*. In contrast to *Mei* and advanced avialans[6], the postorbital process of the jugal is high and involved in the formation of a complete postorbital bar in *Aurornis*. The quadratojugal process is a small posteroventrally directed knob. The triradiate postorbital of *Aurornis* is larger than in *Archaeopteryx*[7], but its frontal process seems proportionally shorter than in *Anchiornis*[2]. The robust lacrimal of *Aurornis* is T-shaped in lateral view, with a

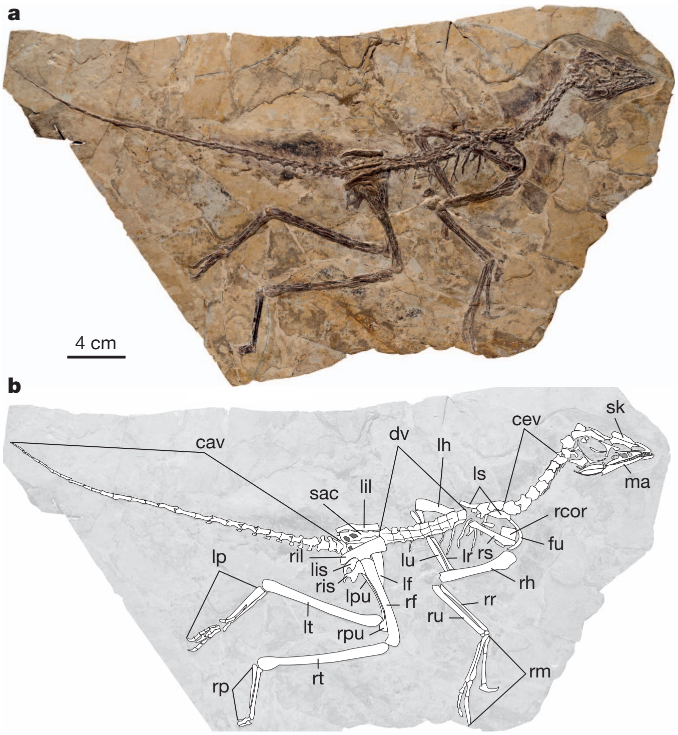


Figure 1 | *Aurornis xui* YFGP-T5198. a, Photograph. **b**, Line drawing. Abbreviations: cav, caudal vertebrae, cev, cervical vertebrae; dv, dorsal vertebrae; fu, furcula; ga, gastralia; lf, left femur; lh, left humerus; lil, left ilium; lis, left ischium; lp, left pes; lpu, left pubis; lr, left radius; ls, left scapula; lt, left tibia; lu, left ulna; ma, mandible; rcor, right coracoid; rf, right femur; rh, right humerus; ril, right ilium; ris, right ischium; rm, right manus; rp, right pes; rpu, right pubis; rr, right radius, rs, right scapula; rt, right tibia; ru, right ulna; sac, sacrum; sk, skull.

[1]Operational Direction 'Earth and History of Life', Royal Belgian Institute of Natural Sciences, rue Vautier 29, 1000 Bruxelles, Belgium. [2]Museo Geologico 'Giovanni Capellini', Via Zamboni 63, I- 40127 Bologna, Italy. [3]Paleontological Institute, Shenyang Normal University, 253 North Huanghe Street, Shenyang 110034, China. [4]Key Laboratory of Vegetation Ecology, Ministry of Education, Northeast Normal University, 5268 Renmin Street, Changchun 130024, China. [5]Eldonia, 9 Avenue des Portes Occitanes, 3800 Gannat, France. [6]Research Center of Paleontology and Stratigraphy, Jilin University, 938 Ximinzhu Street, Changchun, 130021, China. [7]Ocean and Earth Science, National Oceanography Centre, University of Southampton, European Way, Southampton SO14 3ZH, UK.

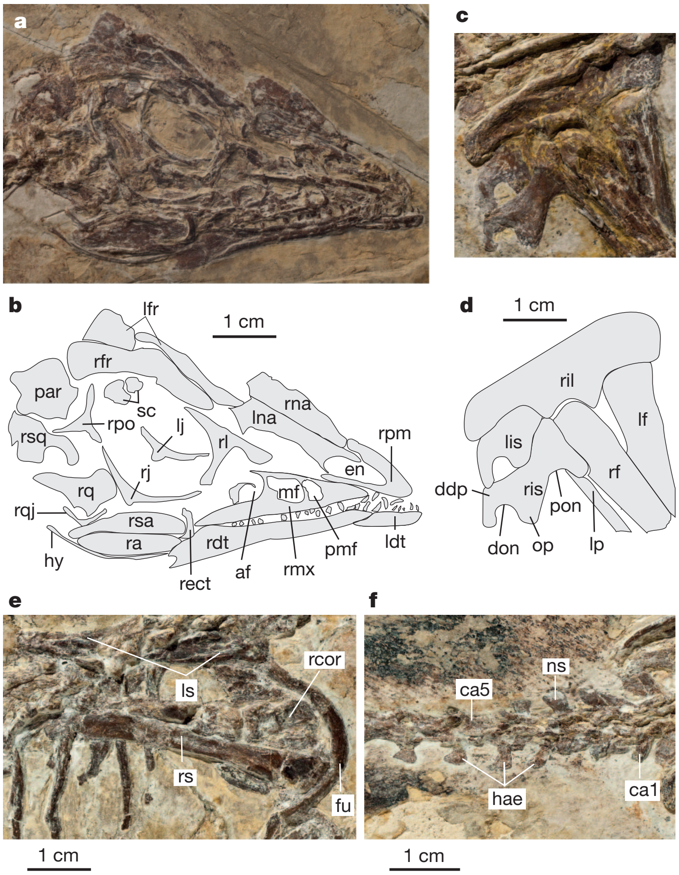


Figure 2 | Selected skeletal elements of *Aurornis xui* YFGP-T5198. **a**, Photograph of skull and mandible in right lateral view. **b**, Line drawing of skull and mandible in right lateral view. **c**, Photograph of pelvis in right lateral view. **d**, Line drawing of pelvis in right lateral view. **e**, Photograph of the scapular girdle. **f**, Photograph of proximal portion of the tail in right lateral view. Abbreviations: af, antorbital fenestra; ca, caudal; ddp, dorsodistal process; don, distal obturator notch; en, external naris; fu, furcula; hae, haemapophyses; hy, hyoid; ldt, left dentary; lf, left femur; lfr, left frontal; lis, left ischium; lj, left jugal; lna, left nasal; lp, left pubis; ls, left scapula; mf, maxillary fenestra; ns, neural spine; op, obturator process; par, parietal; pmf, promaxillary fenestra; pon, proximal obturator notch; ra, right angular; rcor, right coracoid; rdt, right dentary; rect, right ectopterygoid; rf, right femur; rfr, right frontal; ril, right ilium; ris, right ischium; rj, right jugal; rl, right lacrimal; rmx, right maxilla; rna, right nasal; rpm, right premaxilla; rpo, right postorbital; rq, right quadrate; rqj, right quadratojugal; rs, right scapula; rsa, right surangular; rsq, right squamosal; sc, scleral plates.

posterior process that is longer than its anterior process, perpendicular to the descending process and which participates in about half the length of the dorsal margin of the orbit, contrasting with the proportionally shorter posterior process in *Anchiornis*[2], *Archaeopteryx*[3,7] and troodontids[6,8–10], and with the vestigial anterior process in *Eosinopteryx*[4]. The frontal of *Aurornis* is about 45% of total skull length, contrasting with a proportionally shorter element in *Anchiornis*[2]. Unlike in *Anchiornis*[2] and dromaeosaurids[8,11], a paraquadrate notch is not developed on the quadrate. The anterior half of the dentary is more slender than that of *Anchiornis* and has subparallel dorsal and ventral margins. The presence of a posteriorly widening groove on the labial surface of the dentary is a derived feature of *Aurornis* that is also shared with *Anchiornis*[2], *Xiaotingia*[3], *Eosinopteryx*[4], *Archaeopteryx*[3,12] and troodontids[6,9]. The maxillary teeth are tiny, triangular in labial view and unserrated as in *Anchiornis*[2], *Mei*[6] and *Byronosaurus*[13], contrasting with the blunt teeth of *Xiaotingia*[3]. As in *Anchiornis*, the middle and posterior maxillary teeth of *Aurornis* are more sparsely distributed than the anterior ones[2].

Seven postaxial cervical vertebrae are present in *Aurornis*. Shared with *Archaeopteryx*[7], the cervical ribs of *Aurornis* are distinctly longer than their corresponding vertebrae, contrasting with the shorter ribs of *Eosinopteryx*[4] and *Troodon*[9]. The trunk of *Aurornis* is about 30% the length of the hindlimb, similarly proportioned to *Mei*[6], but distinctly shorter (42%) than *Anchiornis*[2]. The neural spines of the middle and posterior dorsals of YFGP-T5198 are particularly shortened. The synsacrum is composed of five vertebrae and the tail of about 30, making the tail in this animal proportionally longer (approximately four times the length of the femur) than in *Mei* (3.17)[6], *Archaeopteryx* (3.27)[7] and *Eosinopteryx* (2.71)[4]. The anteriormost caudals are short. A neural spine is developed on only the anterior third or fourth caudals. The chevrons resemble those of *Archaeopteryx*[7]; between the anteriormost centra, these are vertically oriented rectangular plates but become proportionally lower and develop elongated posterior extensions posteriorly from the seventh to the twelfth caudals (Fig. 2f). *Anchiornis* has hook-like proximal chevrons[1], whereas *Eosinopteryx* is characterized by small rod-like elements that extend below the nine proximal caudals[4].

The scapula of *Aurornis* is slender and perfectly straight (Fig. 2e). The furcula appears more robust than in both *Anchiornis* and *Eosinopteryx*, more resembling the condition in *Archaeopteryx*[7] and *Confuciusornis*[14]. The arm is long, about 80% of leg length, approaching the condition of *Archaeopteryx* (87–104%)[7]. As in *Xiaotingia*[3], the humerus of *Aurornis* is slightly shorter (88%) than the femur, whereas this element is only half the length of the femur in *Mei*[6] and distinctly longer in *Archaeopteryx* (1.12–1.24)[7]. As in *Anchiornis* and *Eosinopteryx*[4], the radius and ulna of *Aurornis* are straight and closely contact each other; in *Xiaotingia*[3], *Archaeopteryx*[7], *Mei*[6], dromaeosaurids[11,15], and *Sinornithoides*[16], the ulna is distinctly bowed distally and is much thicker than the radius. In YFGP-T5198 the manus is slightly longer than the femur (manus/femur lengths =1.09) as in *Eosinopteryx* (1.17)[4], contrasting with the shorter manus in *Mei* (0.82)[6] and the more elongate hands of *Anchiornis* (1.56) (ref. 2) and *Archaeopteryx* (1.4–1.56)[7]. Metacarpal I is about one-third the length of metacarpal II in *Aurornis* (Supplementary Fig. 5a, b); metacarpal III is shorter and more slender than metacarpal II, as in non-scansoriopterygid paravians. The long manual phalanx I-1 (minimum width 3 mm) is more robust than the radius (minimum width 1.5 mm). The dorsal margin of the postacetabular process of the ilium remains subhorizontal along its entire length (Fig. 2c, d), although it is usually oblique ventrally in other known basal paravians[3]. The ischium of *Aurornis* is shortened, less than 30% the length of the femur. A triangular obturator process is present at the mid-point of the ischium; this is proximodistally longer than high, as in *Anchiornis*[1] and *Eosinopteryx*[4], contrasting with the shorter and distally placed obturator process in *Rahonavis*[17]. There is no trace of a proximodorsal process on the ischium, as in Scansiopterygidae[18] and unlike in Unenlagiinae[19], *Rahonavis* and other basal avialans[17]. The distal end of the ischium is dorsoventrally expanded, formed by a long, robust dorsodistal process, and by a shorter and stout hook-like ventrodistal process that distally delimits a distal obturator notch larger than in *Archaeopteryx*[7]. The femur is slightly bowed anteriorly in lateral view and has a prominently developed lesser trochanter. The tibia (137% of femoral length) and pes (111% of the femoral length) are proportionally shorter in *Aurornis* than in *Anchiornis* (respectively 161% and 156%)[2]. Metatarsal I is slender and more elongate than in other known paravians, being about 30% of metatarsus length (Supplementary Fig. 5c, d). Metatarsal III is transversely compressed, suggesting a sub-arctometatarsalian condition. Pedal digit I lies on the medioplantar side of metatarsal II, as in *Archaeopteryx*[7,20], but contrasts with pedal digit I of *Anchiornis* that lies medial to metatarsal II[1]. The phalanges of pedal toes II, III and IV gradually decrease in length proximodistally, as in *Archaeopteryx* and terrestrial cursorial birds[21,22]. Unlike *Anchiornis*[1], the second pedal ungual of YFGP-T5198 is not substantially larger than the others. Traces of plumulaceous feathers, comprising a bundle of filaments joined together proximally and remaining almost parallel as they extend distally, are preserved along the proximal third of the tail, in YFGP-T5198 above the neck and around the chest (Fig. 1). Pennaceous feathers are not preserved.

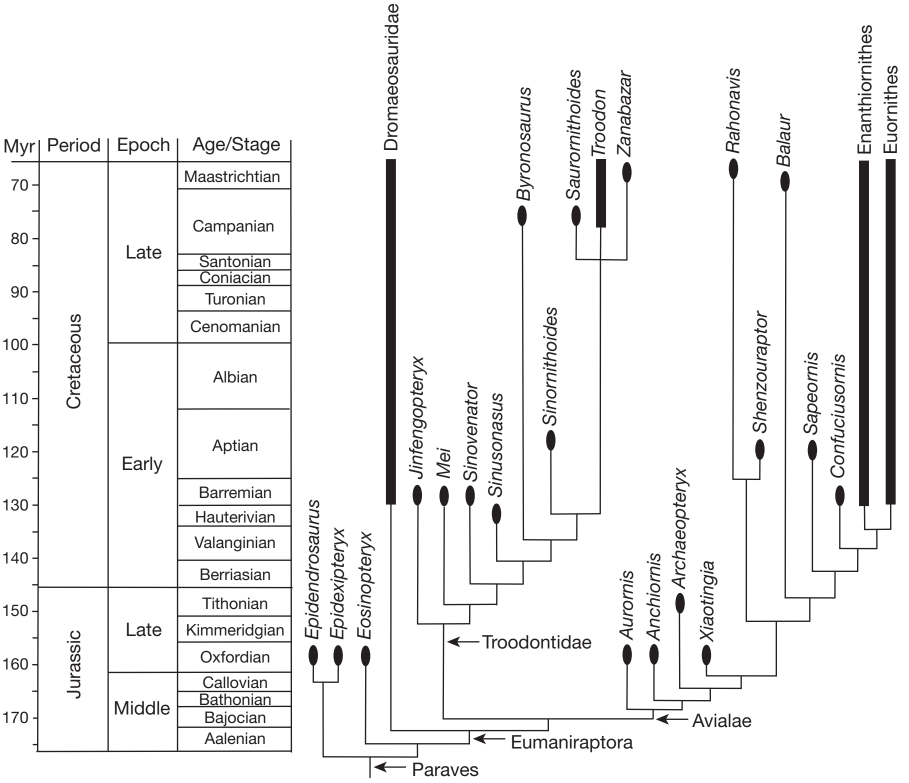


Figure 3 | Phylogenetic relationships of *Aurornis xui* among coelurosaurian theropods. Time-calibrated strict consensus tree of the 216 most-parsimonious trees resulting from our phylogenetic analysis (tree length = 4429; consistency index excluding uninformative characters = 0.27; retention index = 0.58; Supplementary Information). In this hypothesis *Aurornis* is an avialan more basal than *Archaeopteryx*, and Troodontidae is the sister-group of Avialae.

We coded *Aurornis xui* into the largest phylogenetic analysis of basal Paraves so far constructed, including all morphological characters discussed by recent conflicting[2–6,9,10,12,18,19,23] hypotheses (101 taxa versus 992 phylogenetically informative characters, see Supplementary Information). Our result recovers *Aurornis* and *Anchiornis*, both from the Tiaojishan Formation of western Liaoning, as successive basal-most avialans, confirms the Avialan status of *Archaeopteryx* and places Troodontidae as the sister-group for Avialae (Fig. 3). *Epidendrosaurus*, *Epidexipteryx* and *Eosinopteryx*, also from the Middle–Late Jurassic of northeastern China, are here regarded as basal, non-eumaniraptoran paravians. Thus our phylogeny is entirely consistent with the presence of a tetrapterygian condition (= four winged) and elongated rectrices in basal eumaniraptorans. We also postulate a single origin for typical forewing-powered flight, generally inferred to be present only in more derived birds[5,24]; shifting *Archaeopteryx* into deinonychosaurs[3,4] minimally implies two origins (in *Archaeopteryx* and in 'true' birds) or a much more complex situation, with an earlier origin close to the base of Paraves for forewing-driven flight and subsequent modifications to the tetrapterygian condition in various deinonychosaurs[5]. These relationships are also consistent with the recent discovery of potentially four-winged flight surfaces in a range of Mesozoic basal birds[25].

This new comprehensive phylogeny shows that basal avialans (*Aurornis*, *Anchiornis*, *Xiaotingia*) were already diversified in northern China during the Middle–Late Jurassic. These new data, combined with the presence of *Archaeopteryx* in the Tithonian stage of Germany, show that avialans were widespread throughout Eurasia at the end of the Jurassic. In contrast, dromaeosaurids and troodontids are conspicuously absent from Jurassic deposits in Asia, and only a few isolated teeth from the Late Jurassic of Europe have tentatively been identified as belonging to dromaeosaurids[26,27]. Possibly paravian teeth have also been reported from the Middle Jurassic of England[28]. The Jehol Biota of northeastern China testifies to the early diversification of dromaeosaurids (that is, four described genera so far, bearing in mind that many taxa have yet to be formally described), troodontids (also four described genera so far) and the evolutionary explosion of avialans (more than 30 named genera, albeit with nomenclatural problems) at the beginning of the Cretaceous period.

METHODS SUMMARY

This published work and the nomenclatural acts it contains have been registered in ZooBank, the proposed online registration system for the International Code of Zoological Nomenclature. The ZooBank life science identifiers can be resolved and the associated information viewed by appending the life science identifiers to the prefix http://zoobank.org/. The life science identifiers for this publication are urn:lsid:zoobank.org:pub:ACD22438-7DAE-407E-9D79-266D781E1ED2 and urn:lsid:zoobank.org:act:7C240271-6ED2-4633-9597-EF480AC4B811.

Received 8 February; accepted 8 April 2013.
Published online 29 May 2013.

1. Xu, X. *et al.* A new feathered maniraptoran dinosaur fossil that fills a morphological gap in avian origin. *Chin. Sci. Bull.* **54,** 430–435 (2009).
2. Hu, D.-Y., Hou, L., Zhang, L. & Xu, X. A pre-*Archaeopteryx* troodontid theropod with long feathers on the metatarsus. *Nature* **461,** 640–643 (2009).
3. Xu, X., You, H., Du, K. & Han, F. An *Archaeopteryx*-like theropod from China and the origin of Avialae. *Nature* **475,** 465–470 (2011).
4. Godefroit, P. *et al.* Reduced plumage and flight ability of a new paravian theropod from China. *Nature Commun.* **4,** 1394 (2013).
5. Lee, M. S. Y. & Worthy, T. H. Likelihood reinstates *Archaeopteryx* as a primitive bird. *Biol. Lett.* **8,** 299–303 (2012).
6. Xu, X. & Norell, M. A. A new troodontid from China with avian-like sleeping posture. *Nature* **431,** 838–841 (2004).
7. Wellnhofer, P. *Archaeopteryx—Der Urvogel von Solnhofen* (Friedrich Pfeil, 2008).
8. Hwang, S. H., Norell, M. A., Ji, Q. & Gao, K. New specimens of *Microraptor zhaoianus* (Theropoda: Dromaeosauridae) from northeastern China. *Am. Mus. Novit.* **3381,** 1–44 (2002).
9. Makovicky, P. J. & Norell, M. A. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 184–195 (Univ. California Press, 2004).
10. Xu, X. *et al.* A basal troodontid from the Early Cretaceous of China. *Nature* **415,** 780–784 (2002).
11. Norell, M. A. & Makovicky, P. J. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 196–209 (Univ. California Press, 2004).
12. Elzanowski, A. & Wellnhofer, P. Cranial morphology of *Archaeopteryx*: evidence from the seventh skeleton. *J. Vertebr. Paleontol.* **16,** 81–94 (1996).
13. Makovicky, P. J., Norell, M. A., Clark, J. M. & Rowe, T. E. Osteology and relationships of *Byronosaurus jaffei* (Theropoda: Troodontidae). *Am. Mus. Novit.* **3402,** 1–32 (2003).
14. Chiappe, L. M., Ji, A., Ji, Q. & Norell, M. A. Anatomy and systematics of the Confuciusornithidae (Theropoda: Aves) from the late Mesozoic of northeastern China. *Bull. Am. Mus. Nat. Hist.* **242,** 1–89 (1999).
15. Zheng, X. *et al.* A short-armed dromaeosaurid from the Jehol Group of China with implications for early dromaeosaurid evolution. *Proc. R. Soc. B* **277,** 211–217 (2010).

16. Currie, P. J. & Dong, Z.-M. New information on Cretaceous troodontids (Dinosauria, Theropoda) from the People's Republic of China. *Can. J. Earth Sci.* **38,** 1753–1766 (2001).
17. Forster, C. A., Sampson, S. D., Chiappe, L. M. & Krause, D. W. The theropod ancestry of birds: new evidence from the Late Cretaceous of Madagascar. *Science* **279,** 1915–1919 (1998).
18. Zhang, F. *et al.* A bizarre Jurassic maniraptoran from China with elongate ribbon-like feathers. *Nature* **455,** 1105–1108 (2008).
19. Makovicky, P. J., Apesteguía, S. & Agnolín, F. L. The earliest dromaeosaurid theropod from South America. *Nature* **437,** 1007–1011 (2005).
20. Mayr, G., Pohl, B., Hartman, S. & Peters, D. S. The tenth skeletal specimen of *Archaeopteryx. Zool. J. Linn. Soc.* **149,** 97–116 (2007).
21. Xu, X. & Zhang, F. A new maniraptoran dinosaur from China with long feathers on the metatarsus. *Naturwissenschaften* **92,** 173–177 (2005).
22. Hopson, J. A. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 211–235 (Peabody Museum of Natural History, 2001).
23. Turner, A. H. *et al.* A basal dromaeosaurid and size evolution preceding avian flight. *Science* **317,** 1378–1381 (2007).
24. Gauthier, J. & de Queiroz, K. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 7–41 (Peabody Museum of Natural History, 2001).
25. Zheng, X. *et al.* Hind wings in basal birds and the evolution of leg feathers. *Science* **339,** 1309–1312 (2013).
26. Zinke, J. Small theropod teeth from the Upper Jurassic coal mine of Guimarota (Portugal). *Paläontol. Z.* **72,** 179–189 (1998).
27. van der Lubbe, T., Richter, U. & Knötschke, N. Velociraptorine dromaeosaurid teeth from the Kimmeridgian (Late Jurassic) of Germany. *Acta Palaeontol. Pol.* **54,** 401–408 (2009).
28. Evans, S. E. & Milner, A. R. in *In the Shadow of the Dinosaurs. Early Mesozoic Tetrapods* (eds Fraser, N. V. & Sues, H.-D.) 303–321 (Cambridge Univ. Press, 1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Pohl for making fossils available for study, S. Ge for help and comments on the manuscript and T. Hubin for photographs. This study was supported by a grant (BL/36/62) to P.G. from the SPP Politique scientifique (Belgium). D.H. was supported by the Key Laboratory of Evolutionary Systematics of Vertebrates (CAS2011LESV011) and by the National Natural Science Foundation of China (41172026).

Author Contributions P.G., F.E., A.C. and W.W. designed the project. P.G., A.C., G.D. and H.D.-Y. performed the research. P.G., A.C. and G.D. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.G. (Pascal.Godefroit@naturalsciences.be).

doi:10.1038/nature12176

Distinct behavioural and network correlates of two interneuron types in prefrontal cortex

D. Kvitsiani[1]*, S. Ranade[1]*, B. Hangya[1], H. Taniguchi[1]†, J. Z. Huang[1] & A. Kepecs[1]

Neurons in the prefrontal cortex exhibit diverse behavioural correlates[1–4], an observation that has been attributed to cell-type diversity. To link identified neuron types with network and behavioural functions, we recorded from the two largest genetically defined inhibitory interneuron classes, the perisomatically targeting parvalbumin (PV) and the dendritically targeting somatostatin (SOM) neurons[5–8] in anterior cingulate cortex of mice performing a reward foraging task. Here we show that PV and a subtype of SOM neurons form functionally homogeneous populations showing a double dissociation between both their inhibitory effects and behavioural correlates. Out of several events pertaining to behaviour, a subtype of SOM neurons selectively responded at reward approach, whereas PV neurons responded at reward leaving and encoded preceding stay duration. These behavioural correlates of PV and SOM neurons defined a behavioural epoch and a decision variable important for foraging (whether to stay or to leave), a crucial function attributed to the anterior cingulate cortex[9–11]. Furthermore, PV neurons could fire in millisecond synchrony, exerting fast and powerful inhibition on principal cell firing, whereas the inhibitory effect of SOM neurons on firing output was weak and more variable, consistent with the idea that they respectively control the outputs of, and inputs to, principal neurons[12–16]. These results suggest a connection between the circuit-level function of different interneuron types in regulating the flow of information and the behavioural functions served by the cortical circuits. Moreover, these observations bolster the hope that functional response diversity during behaviour can in part be explained by cell-type diversity.

To investigate whether distinct interneuron types can encode specific behavioural variables we recorded the activity of inhibitory neurons expressing parvalbumin and somatostatin markers (Supplementary Fig. 1a). PV basket cells are thought to control the spiking output of pyramidal neurons[12,14], whereas most SOM interneurons, known as Martinotti cells (Supplementary Fig. 1c, d), target distal dendrites, gating the inputs arriving onto pyramidal cells[13,15,17–20]. To target these interneuron types for recordings, we used PV-Cre and SOM-Cre[21,22] driver mouse lines in combination with adeno-associated viruses to deliver channelrhodopsin-2 (ChR2)[23], rendering neurons light sensitive (Supplementary Fig. 1a, b). Miniature microdrives housing 6 movable tetrodes and an optical fibre were implanted in deep layers of the anterior cingulate cortex (ACC) (Fig. 1a and Supplementary Fig. 1e–g). We recorded well-isolated single units ($n = 1{,}339$ from 6 PV-Cre and 6 SOM-Cre mice) and delivered brief pulses (1 ms) of blue light to elicit short-latency action potentials in ChR2-expressing neurons that served as a physiologic tag[24] (Fig. 1b, c). To identify directly light-activated units we developed an optical-tagging test based on a statistical measure that yields a P value testing whether light-activation induced significant changes in spike timing (Fig. 1d and Supplementary Fig. 2, see Methods). Significantly activated units ($P < 0.01$) showed similar spontaneous and light evoked waveforms (correlation coefficient, $r > 0.85$, Fig. 1b and Supplementary Fig. 2c), low-latency light-induced response (< 4 ms), and low first-spike jitter (Fig. 1e, f), signatures of direct light-activation.

Extracellularly recorded units are traditionally classified based on spike width and firing rate, with narrow-spiking and fast-firing neurons categorized putatively as PV interneurons[1,25]. Indeed, most identified PV neurons were narrow-spiking with high firing rates (219 ± 10 μs, 31 ± 3 Hz, $n = 23$, Fig. 1e), whereas the spike-width distribution for SOM units was bimodal (Fig. 1e, bottom): a third of neurons had narrow spikes (‘NS’, <270 μs) and high firing rates (212 ± 7 μs, 16 ± 4 Hz, $n = 13$) and the rest showed markedly wider spike waveforms and lower firing rates (‘WS’, 327 ± 7 μs, 4 ± 1 Hz, $n = 22$).

Having identified PV and SOM interneurons, we first examined their effect on local circuit activity. Synchronous photostimulation of ChR2-expressing PV or SOM neurons had markedly different network effects, with PV neurons imposing brief uniform inhibition on nearby neurons[26], and SOM neurons exerting longer and more variable inhibition (Fig. 2a, b and Supplementary Fig. 3a, b). These differences cannot be accounted for by systematic differences in the number of photo-activated neurons (Supplementary Fig. 4) and indicate that SOM and PV neurons exert distinct inhibitory footprints on network activity.

Optogenetic identification of many individual interneurons, in combination with simultaneous recording from a large number of their neighbours, allowed us to investigate the physiological impact of different inhibitory subtypes during behavioural epochs without light stimulation. To identify possible functional connectivity between neurons, we computed cross-correlograms (CCGs)—counts of spike co-occurrences in the putative pre- and postsynaptic neuron pairs at different time lags[27] (Fig. 2c). Significant short-latency interactions were rare among pairs of unidentified ACC neurons (3.2% inhibitory, 5.2% excitatory, 1.3% both, out of 2,945 pairs, bootstrap test with $P < 0.001$ used for all CCG significance testing). Remarkably, 5/7 pairs of PV neurons showed interactions with 3/7 firing in millisecond zero-lag synchrony, and 4/7 inhibited each other (trough at 2.25 ± 0.5 ms, Fig. 2c and Supplementary Fig. 5a, c). PV neurons also showed a high prevalence of short-timescale correlations with unidentified neurons (38/152 pairs, $P < 0.001$, Fig. 2c and Supplementary Fig. 5c), often with detectible inhibition (trough at 2.39 ± 1.3 ms, 18/152 pairs, $P < 0.001$). These results demonstrate that the PV population is capable of millisecond synchronization with fast and precise inhibitory effect on local neural activity.

In contrast to PV pairs, we found no short-timescale correlations between SOM pairs (0/11, 7 WS–WS and 4 NS–WS pairs, Fig. 2c and Supplementary Fig. 5b), and the influence of both NS-SOM and WS-SOM on unidentified neurons was sparser and more diverse (15/169 pairs, inhibitory in 2/169, $P < 0.001$, Fig. 2c and Supplementary Fig. 5c). The weak observable effect of SOM neurons on the firing output of their neighbours could be due to dendritic inhibition generating input suppression, which is expected to be more difficult to detect using a cross-correlation approach. Thus, PV and SOM interneurons

[1] Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. †Present address: Max Planck Florida Institute for Neuroscience, One Max Planck Way, Jupiter, Florida 33458-2906, USA.

*These authors contributed equally to this work.

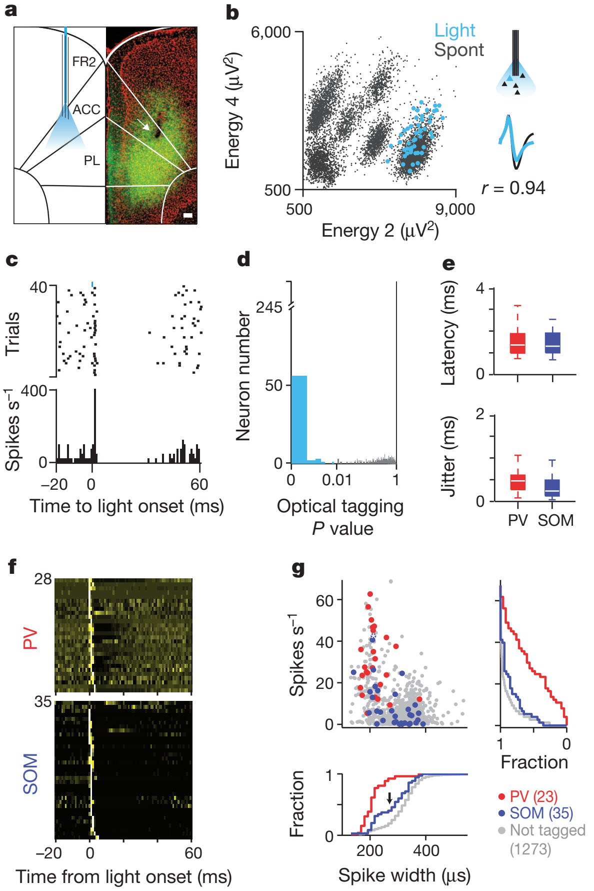


Figure 1 | Optogenetic tagging of genetically-defined interneurons in behaving mice. **a**, Coronal section from a SOM-IRES-Cre mouse (green, ChR2; red, DAPI (4′,6-diamidino-2-phenylindole)). Arrow indicates electrolytic lesion of recording site in the ACC. Scale bar, 100 μm. ACC, anterior cingulate; FR2, frontal region 2; PL, prelimbic cortex. **b**, Spike sorting example. Unclustered spikes plotted in waveform energy space from two tetrode channels (energy 2, energy 4). Light-evoked spikes superimposed in blue. Bottom right, average spontaneous and light-evoked waveforms. **c**, Spike raster (top) and peri-stimulus time histogram (PSTH) (bottom) for the light-activated cell in **b**, aligned to light onset. Light pulse shown in blue (duration, 1 ms; power, ~100 mw per mm^2; frequency, 10 Hz). **d**, Histogram of SALT (stimulus-associated spike latency test) for optical tagging yielded strongly bimodal distribution of *P* values; ($P < 0.01$, blue). **e**, Box plots for all tagged PV (red) and SOM (blue) neurons show low light-evoked first-spike latency (top) and small jitter (bottom). **f**, *z*-scored PSTH of all tagged PV (top) and SOM (bottom) interneurons in response to 1-ms blue light stimulation. **g**, Firing rate as a function of spike width for PV, SOM and not tagged neurons. White asterisk indicates neuron in **b** and **c**. Cumulative histograms of firing rate (top right) and spike width (bottom) are plotted for all groups. Arrow marks mode separation of spike width (NS and WS) distribution for SOM neurons.

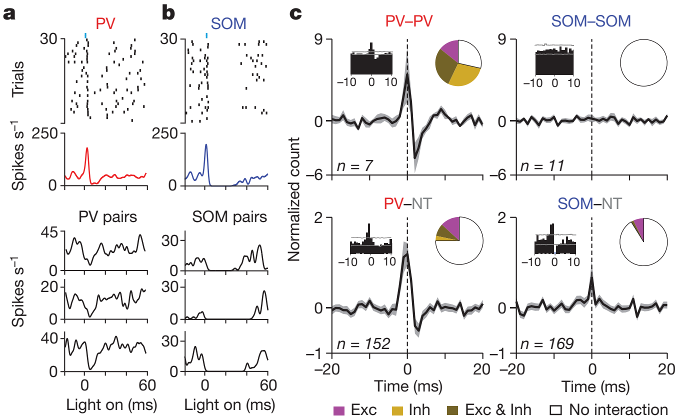


Figure 2 | Distinct inhibitory effect of SOM and PV interneurons. **a**, **b**, Top, spike raster and PSTH of PV (**a**) and SOM (**b**) interneurons aligned to light onset. Bottom, PSTH of three simultaneously recorded unidentified neurons (PV pairs and SOM pairs). **c**, Average cross-correlograms (CCG) between PV–PV (top left), PV–Not tagged (NT) (bottom left), SOM–SOM (top right), and SOM-Not tagged (NT) (bottom right) neuron pairs (shaded area indicates s.e.m.); examples of significant pairwise interactions (left inset) and summary for statistically significant CCG interactions (right inset). Exc, excitatory; Inh, inhibitory.

form distinct inhibitory networks: a fast, synchronous PV network generating strong, transient inhibition and an asynchronous SOM network with weaker effect on firing output.

We explored whether these cell-type differences in network functions are also reflected in specific behavioural correlates. To engage neural ensembles in the ACC we used a task that incorporated cue-based prediction, temporal control of actions and reward foraging decisions (Fig. 3a). Mice were trained to run back and forth on a linear track between two platforms to collect water rewards; entering one platform ('trigger zone') enabled reward availability at the other ('reward zone'). As mice ran back to the reward zone platform, reward size was cued by an auditory signal. This task mimics the self-paced timing of foraging behaviours and exploits the natural tendency of mice to trade staying in a rewarded, safe area with running on an elevated open track to enable future reward collection. Behavioural performance was sensitive to anticipated reward outcomes because on a subset (15%) of trials in which the cue and reward were omitted, mice slowed their speed during reward zone approach (Fig. 3b).

We examined the responses of a population of 1,034 neurons (from 4 PV-Cre and 6 SOM-Cre mice) in the task. Neurons responded at several behavioural events and modulated their firing by different behavioural variables. For instance, as expected of neurons in the ACC[3,9,11], we found single neuron correlates of reward prediction, staying time, and reward outcome and size (Supplementary Fig. 6a). The firing of many individual neurons was selective for single as well as combinations of task variables without any apparent clustering of response properties (Supplementary Fig. 6b–e). Therefore, we used an unbiased approach to determine firing rate modulation patterns for the PV and SOM neurons, which focused our analysis on two behaviourally relevant events, reward approach and leaving (see Methods and Supplementary Fig. 6f). Similar to the example neuron (Fig. 3c), most recorded PV neurons (11/14) phasically increased their firing as mice left the reward zone (Fig. 3d and Supplementary Figs 7c and 8a, b). To test the homogeneity and specificity of these event-related response profiles we used a resampling approach and compared PV interneurons to the unidentified population (see Methods). We found that the temporal response profiles of the PV interneurons were homogeneous ($P < 0.01$, bootstrap test) and distinct ($P < 0.001$) compared to randomly selected groups of neurons (see also Supplementary Fig. 9d). Moreover, knowledge of PV identity carried approximately twice the information about the time course of responses than knowledge that a neuron is narrow-spiking, despite the fact that PV neurons tend to be narrow-spiking ($P < 0.05$, Supplementary Fig. 7b).

The firing of many SOM neurons was strongly suppressed at the time of reward zone entry (13/21, suppression index < 0, $P < 0.01$, permutation test), like the example neuron (Fig. 3e). Similarly, most NS-SOM neurons were suppressed upon entry into the reward zone (9/10, Fig. 3f, bottom, and Supplementary Fig. 8a, b). In contrast, WS-SOM neurons were activated at different moments in time, around the entry into the reward zone (Fig. 3f, top). These profiles were different from both the PV and the unidentified population (Supplementary Fig. 7a, d). Together with their local-circuit effects described above, these observations support the idea that SOM neurons comprise at least two functional subtypes[18,19,28], a narrow-spiking, more

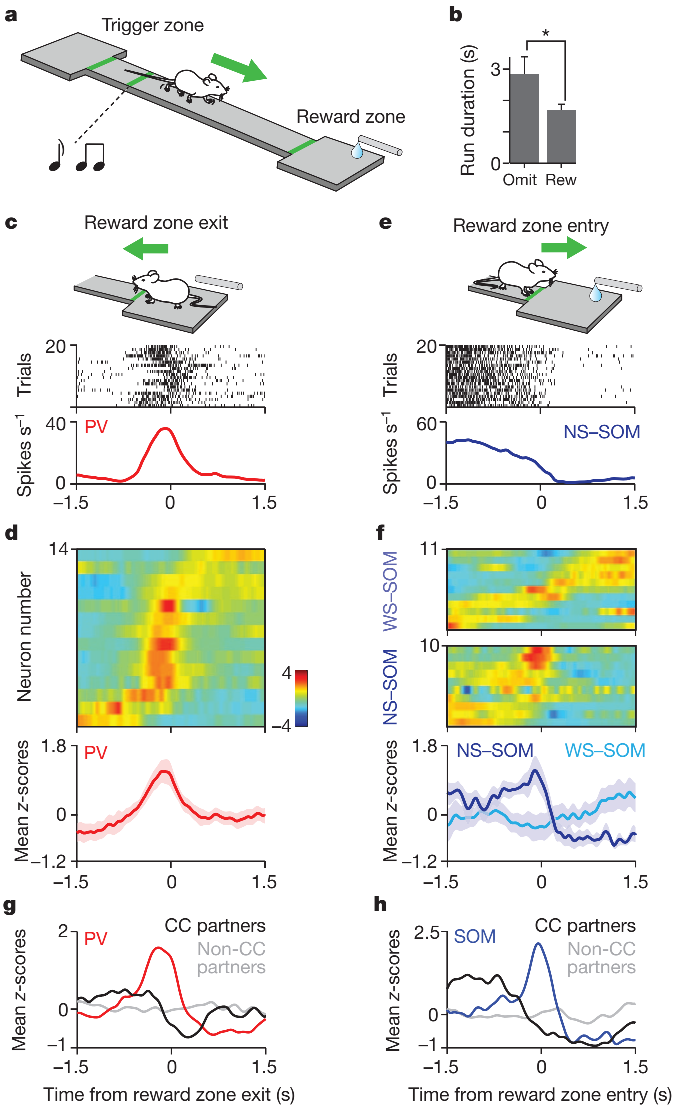


Figure 3 | Distinct behavioural correlates of PV and SOM interneurons. **a**, Cartoon of mouse performing the reward foraging task. **b**, Average run duration for rewarded (Rew) and omission (Omit) trials ($n = 63$ sessions, $P < 0.05$, Mann–Whitney U-test). **c**, Spike raster and peri-event time histogram (PETH) for an identified PV interneuron aligned to time of reward zone exit. **d**, Top, z-scored PETHs of 14 PV neurons sorted by latency to half-peak firing (colors from blue to red indicate low to high normalized firing rate of neurons, respectively); bottom, mean z-scored response (shaded area indicates s.e.m.). **e**, Spike raster and PETH for a NS-SOM interneuron aligned to the time of reward zone entry. **f**, Top, z-scored PETHs of 21 SOM neurons. NS-SOM and WS-SOM neurons are separated. Bottom, mean responses for NS-SOM and WS-SOM neurons (shaded area indicates s.e.m.). **g**, Average PETH for PV interneurons (red, $n = 4$) with significant inhibitory cross-correlations, (CC-partners, black, $n = 5$) and non-CC partners (grey, $n = 76$). **h**, Average PETH for SOM interneurons (blue, $n = 3$), CC-partners (black, $n = 3$) and non-CC partners (grey, $n = 34$).

homogeneously responding population and a wide-spiking population with heterogeneous response profiles.

We examined whether these cell-type-specific differences in behavioural correlates are also reflected in their synaptic partners. We characterized the responses of unidentified wide-spiking (putative pyramidal, pPyr) neurons that showed significant inhibitory cross-correlations with identified interneurons (CC-partners, for example, Fig. 2c, lower panels, inset). Notably, PV→pPyr and SOM→pPyr pairs, which both showed negative cross-correlations on the time scale of milliseconds, exhibited opposing behavioural response-dynamics on the time scale of seconds (Fig. 3g, h). This was not observed for simultaneously recorded neurons without significant short-term interactions with PV or SOM neurons (Fig. 3g, h). These results reveal that functional connectivity, as identified by millisecond cross-correlations indicative of anatomical connections, also predicts post-synaptic neural responses on the time scale of seconds as relevant for behaviour.

Finally, we sought to better understand the behavioural functions of the PV population. We wondered whether the phasic recruitment of PV neurons is related to a specific movement or reflects a more abstract behavioural variable (Supplementary Fig. 9c). Specifically, the ACC has been implicated in foraging decisions[9,11]—whether to stay or to leave. Therefore, we trained mice on a task version in which they were rewarded at a water port after a fixed 1 s delay from entry. In this task variant, the motor program required for the leaving action was a backward movement (Fig. 4a, cartoon), distinct from the forward movement corresponding to the reward zone exit in the original task version. In addition, this enabled more precise measurements of behavioural timing. Mice stayed inside the port for varying durations (2.0 ± 2.2 s) to consume water reward then exited to initiate the next trial. We found that PV neurons responded with a large phasic firing rate elevation around the time of exit from the reward port (Fig. 4d, Supplementary Fig. 8a, b; 11/12 neurons with activation index > 0, $P < 0.05$, permutation test). Because mice could freely exit at any time, we wondered if the activity of these neurons was modulated by the duration of their stay inside the reward port. Indeed, we observed that the firing rate of PV neurons parametrically increased with longer staying times on a trial-by-trial basis (Fig. 4a–d). A similar representation of stay duration has been found in monkey ACC during a foraging task, which was shown to signal the negative value of staying or equivalently the likelihood of leaving during foraging decisions[9]. This suggests that the graded phasic response of PV neurons in ACC is related to a foraging decision, to leave the reward consumption area and initiate a new run.

Our findings demonstrate that two major classes of interneurons not only provide distinct modes of inhibition but also display unique behavioural correlates, with temporal and functional specificity comparable to principal neurons. Out of the many behavioural events in the task, the homogeneous responses of PV and NS-SOM interneurons bracketed a defined epoch: from reward approach to leaving, and represented a specific behavioural variable, staying time at the reward zone, critical for foraging decisions, a central function attributed to ACC[9,10]. How can this temporal and behavioural specificity be

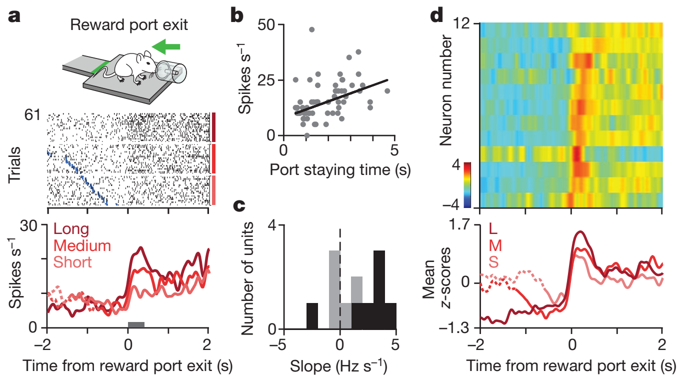


Figure 4 | PV interneurons in the ACC signal stay duration at foraging decisions. **a**, Mouse exiting the reward port (inset). Response of a PV neuron during reward port exit. Raster is sorted by staying time in the port and grouped into terciles. Blue ticks denote water valve offset. PETH is shown for each tercile. **b**, Linear regression between firing rate of the neuron in **a**, (epoch indicated by a grey bar) and staying time is significantly positive ($r = 0.16$, slope, 3.63, $P < 0.005$). **c**, Histogram of regression slopes for all PV neurons. Black bars indicate significant ($P < 0.05$) regression. **d**, Top, z-scored PETHs of 12 PV neurons aligned to reward port exit sorted according to latency of half-peak firing. Bottom, average PETH for PV population grouped into staying time terciles. L, M and S denote long-, medium- and short-staying times, respectively.

understood in the context of our current knowledge of interneurons? First, tuning specificity may arise from the dense, convergent local input these interneuron types receive[7,29], enabling them to 'summarize' local neural activity, which may be particularly high at the moments when a region is engaged in a task[30]. Second, PV interneurons have been implicated in controlling pyramidal cell output[12,14,16], consistent with the synchronous firing and strong inhibitory coupling we observed. In contrast, SOM neurons are thought to gate long-range inputs to principal cells[13,15,17,20], and their asynchronous activation and weaker inhibitory impact on firing output is consistent with this role. In our behaviour, input and output regulation might be expected around the foraging decision, consistent with the observed suppression of NS-SOM interneurons during approach followed by the activation of PV interneurons at reward exit. Taken together, our findings suggest a conceptual model in which these interneuron subtypes specialize in temporally regulating the flow of information in a given cortical circuit during the behavioural events relevant to that area. In summary, these observations bolster the long-held hope that probing identified cell-types will reveal the intrinsic logic of cortical circuits under more natural behavioural settings[5,6].

METHODS SUMMARY

All procedures involving animals were carried out in accordance with National Institutes of Health standards as approved by the Cold Spring Harbor Laboratory Institutional Animal Care and Use Committee.

Full Methods and any associated references are available in the online version of the paper.

Received 27 August 2012; accepted 10 April 2013.

Published online 26 May 2013.

1. Constantinidis, C., Williams, G. V. & Goldman-Rakic, P. S. A role for inhibition in shaping the temporal flow of information in prefrontal cortex. *Nature Neurosci.* **5,** 175–180 (2002).
2. Machens, C. K., Romo, R. & Brody, C. D. Functional, but not anatomical, separation of ''what'' and ''when'' in prefrontal cortex. *J. Neurosci.* **30,** 350–360 (2010).
3. Narayanan, N. S. & Laubach, M. Top-down control of motor cortex ensembles by dorsomedial prefrontal cortex. *Neuron* **52,** 921–931 (2006).
4. Wallis, J. D. & Kennerley, S. W. Heterogeneous reward signals in prefrontal cortex. *Curr. Opin. Neurobiol.* **20,** 191–198 (2010).
5. Isaacson, J. S. & Scanziani, M. How inhibition shapes cortical activity. *Neuron* **72,** 231–243 (2011).
6. Klausberger, T. *et al.* Brain-state- and cell-type-specific firing of hippocampal interneurons *in vivo. Nature* **421,** 844–848 (2003).
7. Markram, H. *et al.* Interneurons of the neocortical inhibitory system. *Nature Rev. Neurosci.* **5,** 793–807 (2004).
8. Hartwich, K., Pollak, T. & Klausberger, T. Distinct firing patterns of identified basket and dendrite-targeting interneurons in the prefrontal cortex during hippocampal theta and local spindle oscillations. *J. Neurosci.* **29,** 9563–9574 (2009).
9. Hayden, B. Y., Pearson, J. M. & Platt, M. L. Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neurosci.* **14,** 933–939 (2011).
10. Kolling, N., Behrens, T. E., Mars, R. B. & Rushworth, M. F. Neural mechanisms of foraging. *Science* **336,** 95–98 (2012).
11. Quilodran, R., Rothe, M. & Procyk, E. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron* **57,** 314–325 (2008).
12. Atallah, B. V., Bruns, W., Carandini, M. & Scanziani, M. Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron* **73,** 159–170 (2012).
13. Gentet, L. J. *et al.* Unique functional properties of somatostatin-expressing GABAergic neurons in mouse barrel cortex. *Nature Neurosci.* **15,** 607–612 (2012).
14. Lovett-Barron, M. *et al.* Regulation of neuronal input transformations by tunable dendritic inhibition. *Nature Neurosci.* **15,** 423–430 (2012).
15. Murayama, M. *et al.* Dendritic encoding of sensory stimuli controlled by deep cortical interneurons. *Nature* **457,** 1137–1141 (2009).
16. Royer, S. *et al.* Control of timing, rate and bursts of hippocampal place cells by dendritic and somatic inhibition. *Nature Neurosci.* **15,** 769–775 (2012).
17. Kapfer, C., Glickfeld, L. L., Atallah, B. V. & Scanziani, M. Supralinear increase of recurrent inhibition during sparse activity in the somatosensory cortex. *Nature Neurosci.* **10,** 743–753 (2007).
18. Ma, Y., Hu, H., Berrebi, A. S., Mathers, P. H. & Agmon, A. Distinct subtypes of somatostatin-containing neocortical interneurons revealed in transgenic mice. *J. Neurosci.* **26,** 5069–5082 (2006).
19. McGarry, L. M. *et al.* Quantitative classification of somatostatin-positive neocortical interneurons identifies three interneuron subtypes. *Front. Neural Circuits* **4,** 12 (2010).
20. Silberberg, G. & Markram, H. Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells. *Neuron* **53,** 735–746 (2007).
21. Hippenmeyer, S. *et al.* A developmental switch in the response of DRG neurons to ETS transcription factor signaling. *PLoS Biol.* **3,** e159 (2005).
22. Taniguchi, H. *et al.* A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron* **71,** 995–1013 (2011).
23. Sohal, V. S., Zhang, F., Yizhar, O. & Deisseroth, K. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature* **459,** 698–702 (2009).
24. Lima, S. Q., Hromádka, T., Znamenskiy, P. & Zador, A. M. PINP: a new method of tagging neuronal populations for identification during *in vivo* electrophysiological recording. *PLoS ONE* **4,** e6099 (2009).
25. Csicsvari, J., Hirase, H., Czurkó, A., Mamiya, A. & Buzsáki, G. Fast network oscillations in the hippocampal CA1 region of the behaving rat. *J. Neurosci.* **19,** RC20 (1999).
26. Cardin, J. A. *et al.* Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* **459,** 663–667 (2009).
27. Fujisawa, S., Amarasingham, A., Harrison, M. T. & Buzsáki, G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neurosci.* **11,** 823–833 (2008).
28. Xu, X. & Callaway, E. M. Laminar specificity of functional input to distinct types of inhibitory cortical neurons. *J. Neurosci.* **29,** 70–85 (2009).
29. Ascoli, G. A. *et al.* Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nature Rev. Neurosci.* **9,** 557–568 (2008).
30. Isomura, Y., Harukuni, R., Takekawa, T., Aizawa, H. & Fukai, T. Microcircuitry coordination of cortical motor information in self-initiation of voluntary movements. *Nature Neurosci.* **12,** 1586–1593 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the Klingenstein, John Merck, Sloan and Whitehall Foundations to A.K. and the National Institute of Neurological Disorders and Stroke (National Institutes of Health) grant R01NS075531. B.H. received support from the Swartz Foundation and Marie Curie International Outgoing Fellowship within the EU Seventh Framework Programme for Research and Technological Development (2007-2013). D.K. received support from The Robert Lee and Clara Guthrie Patterson Trust Postdoctoral Fellowship and Human Frontier Science Program (2008–2011). We are grateful to K. Deisseroth, E. Boyden, A. Reid and A. Zador for constructs, B. Burbach and R. Eifert for technical assistance, and to J. Lisman, B. Mensh, S. Shea and A. Zador for comments and discussions.

Author Contributions D.K., S.R. and A.K designed experiments. D.K. and S.R. set up and performed experiments. B.H. developed the optical tagging index. D.K., S.R., B.H. and A.K. analysed data and wrote the paper. H.T. and J.Z.H. generated SOM-Cre mice, discussed results and edited the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.K. (kepecs@cshl.edu).

METHODS

Microdrive construction. We designed two microdrive models that enabled concurrent optical stimulation and recording of neuronal activity. The plastic frame for the drives was designed using AutoCAD Inventor (Autodesk) and 3D printed (Vista Technologies). In one microdrive model (3.5 g) the frame can house up to 10 individually moveable shuttles and is well suited to record and optically activate large population of neurons. Each shuttle has precision holes drilled to attach the tetrode and/or optical ferrule and a miniature screw (0.6 mm outer diameter, 12 mm length) with a pitch of 160 μm. An electronic interface board (EIB32, Neuralynx) connects tetrodes to the preamplifier (HS36, Neuralynx). The fibre-optic probe for optical stimulation consists of a polyimide coated multimode fibre (60 μm diameter, Polymicro Technologies) glued into a fibre-optic ferrule (LC ferrule 80 μm, Precision Fibre Products). The ferrule end is polished using standard optical methods for efficient light coupling and the other end is precisely cleaved for insertion into the brain. The other microdrive model can house up to 5 independently adjustable shuttles and weighs 2.2 g when loaded with a single shuttle driving a bundle of 6 tetrodes and an optical fibre (Supplementary Fig. 1e).

Viral injection. Adeno-associated virus (AAV) 2/9 serotype (8×10^{12} genome copies per ml, UNC Vector Core Facility) carrying EF1a-DIO-ChR2-EYFP or EF1a-DIO-Arch-EYFP construct[23] was injected into 1-month-old PV-Cre and SOM-Cre male mice. Mice were anaesthetized with an intraperitoneal injection of ketamine-xylazine mixture (0.1 mg per gram body weight ketamine, 0.01 mg per gram body weight xylazine). Exposed skin surfaces were occasionally irrigated with lidocaine. A small craniotomy was made above the left dorsomedial prefrontal cortex (2 mm anteroposterior, 0.5 mm mediolateral, from bregma). Virus was injected with a glass micropipette using a Picospritzer (General Valve). Pulses (20–60) of 10 ms duration were delivered at 0.2 Hz starting from a depth of 1.9 mm from the brain surface up to 1.2 mm in 100 μm steps, waiting a minimum of 2–3 min per site to allow diffusion of the virus. Animals were allowed to recover for at least 2 weeks for optimal viral expression.

Microdrive implantation. After anaesthesia a ~1-mm diameter hole was drilled through the skull at the site of viral injection. Animals received supplementary dose of anaesthetic at 30–90-min intervals to maintain depth of anaesthesia. The microdrive was positioned with the help of a stereotaxic arm (David Kopf Instruments) above the craniotomy with protruding tetrodes. The optical fibre and tetrodes were gradually lowered to a depth of 500 μm from the brain surface. Two 0.25-mm diameter stainless steel wires (Alpha Wire Company) were stripped at the end and inserted into cerebellum and right parietal lobe to a depth of ~1 mm below dura to serve as reference and ground electrodes respectively. Two miniature watch screws (Micro-Mark) were fixed into the parietal plates as anchors. The microdrive was secured to the skull with ultraviolet light curable dental cement (Vitrebond Plus) followed by a layer of black dental acrylic (Lang Dental). Tetrodes and optical fibre were lowered by a further 320 μm before mice recovered from anaesthesia. For post-operative analgesia, ketoprofen (2 mg per kilogram body weight) was administered intraperitoneally. Mice were allowed to recover for at least a week.

Foraging task, behavioural setup, training. The behavioural setup consists of an elevated linear track (length 45 cm, width 5 cm) that connects two 8 × 10 cm platforms (termed 'reward zone' and 'trigger zone'). Water rewards were delivered at the reward zone either through a lick tube at the end of the reward zone (Fig. 3a) or a water port designed to precisely monitor timing of port entry and exit (Fig. 4a, inset). Position of the mice on the track was monitored at 30 Hz with a spatial resolution of 2 mm pixel^{-1} using a video tracker that tracked one red and one green light-emitting diode (LED) integrated into the preamplifier on the microdrive (Neuralynx). The behavioural hardware (valves, light-sensors) and the laser were triggered through a data acquisition board (National Instruments PCI-MIO-16E-1) controlled by custom MATLAB programs (MathWorks). Position was tracked by Cheetah recording software (Neuralynx). Owing to the additional weight of the preamplifier and cable we used a custom-designed commutator and counterbalance assembly to enable mice to run more freely. The counterbalance consists of a 40-cm boom moving freely on air bearings with a spherical socket at the end acting as an air-bearing commutator. Precisely controlled and frictionless counterbalancing force was achieved using a pneumatic actuator. The tether was suspended by a hollow ball glued to it that floated inside the socket, and was connected to a slip-ring commutator (PSR-27, Neuralynx) to release torque accumulated by the tether.

One to three weeks after surgery, mice were trained to run on the track. In the initial phase of training, mice were provided with water whenever they approached either the reward zone or poked into the water port. After consuming 20–40 rewards, mice were conditioned to obtain water by running to the opposite end of the platform, the trigger zone, and running back to collect the reward. Entry into the trigger zone activated an auditory cue, which signalled availability of water at the reward zone. Once mice performed about 60 runs we introduced different reward sizes (small: 2–4 μl, large: 6–12 μl) signalled by distinct auditory cues (mixture of high frequency and low frequency tones, 0.1 s duration). Mice also received a reminder auditory cue immediately after exiting the trigger zone (Fig. 3a). On a small fraction of trials (15%) reward was omitted. Mice performed 60–200 trials per session lasting 1–2 h. Animals were kept on a water restriction schedule to maintain 85–90% of free-drinking weight.

Recording and and optical stimulation of genetically identified interneurons. Electrophysiological recordings were performed using a Neuralynx Cheetah 32 system. Electrical signals were split and separately amplified and filtered for local field potentials (LFPs) and single unit activity. The signal was band-pass filtered between 600–6,000 Hz and sampled at 32 kHz to record spiking activity, while LFPs were filtered between 0.1–400 Hz and acquired at 3 kHz. We used 6 tetrodes and one optical fibre to record a total of 1,339 single units from 12 mice. Of these, 1,034 neurons (from 4 PV-Cre and 6 SOM-Cre mice) were recorded during the foraging task and 305 neurons (from 2 PV-Cre mice) were recorded in the port variant of the same task. We recorded a total of 28 PV cells from ACC in 5 PV-Cre animals (5, 5, 6, 2 and 10 cells from each animal out of 15, 14, 19, 24 and 41 sessions, respectively) and 35 cells from 6 SOM-Cre animals (14, 4, 2, 3, 10 and 2 cells per animal with 29, 12, 16, 17, 13 and 12 sessions, respectively). In addition we recorded one PV-Cre animal that gave no tagged neurons. Neurons that had baseline firing rate <1 Hz or showed no activity during perievent periods (window size was specific for each event, see below) were excluded from behavioural analyses.

An optical multimode fibre (55 μm diameter NA = 0.7, Polymicro Technologies) was coupled via a modified LC–LC type connector to a multimode fibre (126 μm diameter, numerical aperture = 0.27, CablesPlus USA), which collected light from a blue laser (473 nm; 20 mW; CrystaLaser). Maximal power at the tip of the fibre ranged from 10% to 30% of power at the light source resulting in 2–6 mW of total output at the fibre tip.

To evaluate the spatial extent of light on brain tissue we conducted (1) photobleaching experiments to measure the area with bleached fluorophore and (2) c-Fos staining around the fibre tip. For photobleaching experiments, blue light (473 nm, 2–4 mW power) was applied continuously for 1 h, whereas for c-Fos induced expression we applied the same light for 1 h in 1 ms pulses at 20 Hz. The spread of light in photobleaching experiments was ~1,000 μm (dorso-ventral axis) by 500 μm (medio-lateral axis). Maximum c-Fos induction occurred within a 0.5 mm^2 area (Supplementary Fig. 4a–c). Because our tetrodes were well within 500 μm from the tip of the optical fibre, light reach is not expected to be a limiting factor for optical tagging.

To avoid a photo-electric artefact due to light stimulation[32], we positioned our tetrodes parallel to the fibre and in cases where we saw an artefact, we minimized it by lowering the light intensity. We verified the validity of optical tagging by comparing the average peak-aligned spontaneous waveform with average light evoked waveform using Pearson's correlation coefficient ($r > 0.85$).

The light stimulation protocol (15–30 min) for optogenetic tagging was performed at the end of each recording session consisting of 1–2 ms light pulses at 4, 10, 16, 40 and 100 Hz frequencies. The fibre and tetrodes were lowered 20–40 μm every day after each recording session. At the end of the experiments, electrolytic lesions were made through individual leads of each tetrode on which a tagged neuron was recorded. We only included optically tagged neurons that were mapped to the anterior cingulate cortex based on the cytoarchitectonic structure of the prefrontal cortex[33].

To reveal the morphology of SOM interneurons we used MADM[34] to visualize the arborisation of these interneurons at single cell resolution.

Data analysis. All data analysis was carried out using built-in and custom-built software in MATLAB (MathWorks).

Spike sorting. Spikes were manually sorted into clusters (presumptive neurons) off-line based on peak amplitude and waveform energy using MClust software (A.D. Redish). Cluster quality was quantified using isolation distance[35] and L-ratio[36]. Clusters with isolation distance <18 or L-ratio >0.2 were excluded (median isolation distance, 29; median L-ratio, 0.033, see Supplementary Fig. 2a, b). Autocorrelation functions were inspected for all putative cells. In cases in which the autocorrelation showed absolute refractory period violations, we improved cluster separation, otherwise, the cluster was excluded.

SALT. We developed a statistical test to determine whether optogenetic activation caused a significant change in the timing of spikes after stimulation onset (Supplementary Fig. 11). The distribution of first spike latencies relative to the light pulse, assessed in a 10 ms window after light-stimulation, was compared to epochs of the same duration in the stimulus-free baseline period. The choice of a 10 ms window size provided sufficient statistical power without limiting the number of detected neurons. To measure the distance between these distributions, we used an information theoretic measure (modified Jensen–Shannon divergence)[37]. Using this metric, we tested the hypothesis that the post-stimulus spike-latency distribution is different from a set of baseline distributions for low frequency light stimulation (4 or 10 Hz) yielding a P-value for significant short-latency light-activation. (See Supplementary Notes for a detailed description and http://kepecslab.cshl.edu/software/ for MATLAB implementation). Note that we also employed a spike shape correlation measure to ensure that our spike sorting was not compromised due to high laser intensities[38]. This is a complementary test as it pertains to spike sorting, whereas SALT tests light effects assuming that spike sorting is correct.

Detection of light-induced inhibition. To detect light-induced inhibition we first determined the putative suppression period using an adaptive smoothing procedure and then evaluated the statistical significance of the firing rate suppression compared to a stimulus-free baseline. First, spike rasters were convolved with a variable kernel Gaussian function to provide a spike density function (SDF) estimate. The kernel width of the Gaussian window was adapted to the local estimate of spiking probability to implement stronger smoothing when information was sparse. Variance was mapped onto spiking probability between 0 (moving average, corresponding to probability of 0) and infinity (Dirac-delta, corresponding to probability of 1). Next, minimal firing was assessed as the minimum of the SDF within 100 ms from light pulse onset. The baseline firing rate was calculated from mean firing probability within a 100 ms window before the start of a pulse train. If the minimal firing after stimulation was lower than 50% of baseline firing rate, then we determined the putative suppression period as the epoch between the half-baseline crossings before and after the minimum. The statistical significance of this suppression was determined by comparing the spike count distribution within this suppression period with an equivalent baseline period using the Mann–Whitney U-test ($P < 0.05$). Note that we used a 50% baseline minimum to provide sufficient statistical power to the spike rate comparison and to avoid false detection of random fluctuations in firing rate.

Cross-correlation analysis. Cross-correlations between spike trains were calculated using 1-ms bin size and their statistical significance was evaluated using a modified temporal shuffling method. To infer putative monosynaptic interactions from extracellularly recorded neurons it is critical to rule out co-firing arising from slow time-scale covariations, for instance due to common input[39,40] or oscillatory modulations[41]. Under the assumption that spike trains are independent of one another, the shift predictor can be used to establish the expected level of firing co-occurrence. However, common input or other slow-time scale fluctuations can create trial-to-trial co-variations independent of synaptic interactions. We dealt with the issue of multiple time-scale effects in two ways. First, we used spectral filtering to remove slow time-scale interactions for which shuffle techniques are ill-suited. Second, we next used temporal shuffling to determine the expected level of correlations.

First, the full-length cross-correlation function was computed and high-pass filtered at 4 Hz. For the shuffling, we pseudo-randomly selected 5,000 instances of 30 ms windows from the filtered cross-correlation function, between 100 ms and 5 s time lags. This is equivalent to calculating the cross-correlation of time-shifted data (sometimes called the shuffling method[25,42,43]). The cross-correlation function was then low-pass filtered at 4 Hz to calculate the slow trends previously filtered out (see above). This slow modulation was added back to the shuffled cross-correlations to obtain estimates of cross-correlation that are not distorted by the filtering procedure. This step is necessary to make the shuffled and the original cross-correlations comparable. Significance limits (set to 0.001 for these analyses) were computed based on the distribution of the shuffled cross-correlations. Statistically significant short-latency suppression after a presynaptic spike is generally taken as evidence for monosynaptic inhibitory connections[27,44] and our results examining identified inhibitory neurons bolster this inference.

For group averages, cross-correlations were standardized by subtracting the mean and dividing by the standard deviation of the shuffled cross-correlograms. Pairwise cross-correlation was performed only on units recorded from different tetrodes to avoid artefactual dips ($0 \pm 750\,\mu$s) in cross-correlogram due to the censored period (spike detection dead time), imposed by spike triggering. For PV and SOM 'CC-partners' (wide spiking neurons with significant inhibitory cross-correlations with PV and SOM interneurons) we also included units from the same tetrodes. Inhibition onto CC partners was considered significant only when dips in cross-correlogram were significant beyond +1 ms for two or more bins. We included spikes collected during behavioural sessions and excluded spikes occurring during the optical tagging epochs.

Identification of putative PV interneurons for cross-correlation analysis. We identified putative PV neurons (pPV) based on three criteria: response profile similarity to identified PV neurons, high firing rate (>15 Hz) and narrow spike width (<270 μs). The combination of these three features enabled us to identify a distinct cluster of pPV neurons. Note that our selection algorithm differs from previous studies in that we could make use of the homogeneous firing pattern of identified PV expressing interneurons with respect to behavioural events (Supplementary Fig. 12a). We only selected putative PV interneurons from the second experiment (Fig. 4) incorporating the water port for precise measurement of exit time, that showed the most homogeneity, thus enabling us to confidently isolate putative PV cells. Also note that our aim was not to find all PV neurons, which appeared as false negatives in our data set due to insufficient ChR2 expression or limited light power, but rather to find some cells that resemble identified PV cells enough to conclude with high confidence that they belong to the PV group. We identified 11 putative PV interneurons. Of these neurons, 3 pairs were recorded simultaneously and 2 were recorded along with identified PV cells, yielding 5 new pairs altogether. We found that 4 out of 5 pairs showed significant short latency interaction in cross-correlograms (3 pairs showed both inhibition and synchrony, 1 pair showed synchrony only) (Supplementary Fig. 12b).

Response modulation index and Gap statistics. In order to quantify and compare the selectivity of neural responses to behavioural events, PETHs (± 0.5-s window, 50 ms resolution) were calculated for all neurons with reference to each event (reward zone exit, reward zone entry, trigger zone exit, trigger zone entry, water valve on). Significance limits were assessed by upper and lower 0.005 percentiles of shuffled PETHs. Shuffling was performed with a similar method to cross-correlation analysis, with random shifts between the firing rates and the events ranging from 10 to 30 s, shuffling was performed 2,000 times). Response modulation index for each neuron and event was computed as the standard deviation of the PETH. To compute overall selectivity profiles, modulation indices for significantly modulated PETHs was averaged for different neuronal populations (PV, SOM and not tagged) for each event (Supplementary Fig. 6f).

Hierarchical clustering was performed on response profiles with respect to three behaviourally relevant events (reward zone entry, reward zone exit and cue presentation) using squared Euclidean distance measure, averaged over the three events. Number of clusters was iterated ranging from 1 to 100 and the gap statistic[45] was calculated to assess the number of clusters naturally present in the data set (see full description of this in Supplementary Notes).

Activation and suppression indices. The modulation indices (activation index for reward zone exit and suppression index for the reward zone entry events) were both calculated using receiver operating characteristic (ROC) analysis to provide a graded measure and a significance value associated with them[46]. These indices represent scaled version of ROC area (AUC) between two firing rate distributions before and after the event (window size, 0.4 s). We scaled the AUC so that it ranges from −1 to 1 with the sign denoting whether a neuron is activated or suppressed.

Modulation index = 2 $(\mathrm{ROC_{area}} - 0.5)$ and $\mathrm{ROC_{area}} = \int_{-\infty}^{\infty} P(f_{\mathrm{before}} = f)P(f_{\mathrm{after}} < f)df$ in which f_{before} and f_{after} refer to the firing rates before and after the relevant event. Statistical significance was evaluated using a permutation test, in which trial order was pseudo-randomly shuffled 1,000 times to yield a P value.

Preference index. To compute preference index during various behavioural epochs, trials were divided into two groups according to cue (cue 1 and cue 2), staying time (shorter and longer than median staying time) and reward size (small and large reward). Firing rates within a fixed peri-event time window (200 ms for cue, 1 s for staying time, and 500 ms for reward preference) were compared using ROC analysis identical to activation and suppression indices. A significant cue preference index of less than 0 means that the neuron preferentially fires for cue 1, whereas more than 0 means preference for cue 2, similarly for staying time preference (long = −1, short = 1), and reward preference (small = −1, large = 1).

Staying time modulation of firing rates. We assessed the dependence of firing rate at the reward port on staying time after exit from the reward port on a trial-by-trial basis using robust regression. To determine activation for PV neurons, a window of 0–0.4 s was used.

Bootstrap test of homogeneity and firing rate modulation. We performed the following bootstrap tests to compare PV PETHs aligned to reward zone exit to the not tagged population.

For the test of homogeneity, the within-group homogeneity of the PV population was computed by averaging pair-wise correlations (Pearson correlation coefficient) of z-scored PETHs aligned to the reward zone-exit event. This estimate of homogeneity was then tested against the average pair-wise correlation calculated for randomly selected groups of not tagged neurons with the same sample size (1,000 bootstrap samples).

For the test of firing rate modulation, the phasic positive modulation of PV neurons was quantified by the average correlation between PV PETHs aligned to reward zone exit and a template of event-locked firing rate increase. This template was computed as mean z-scored PETH of all positively modulated not tagged neurons ($P < 0.05$, activation index > 0, $n = 107$, permutation test). This estimate of positive firing rate modulation for PV neurons was tested against a bootstrap sample of similar estimates for not tagged neurons in the same manner as for the 'test of homogeneity' (see above).

31. Chow, B. Y. *et al.* High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature* **463,** 98–102 (2010).
32. Han, X. *et al.* Millisecond-timescale optical control of neural dynamics in the nonhuman primate brain. *Neuron* **62,** 191–198 (2009).

33. Van De Werd, H. J. J. M., Rajkowska, G., Evers, P. & Uylings, H. B. Cytoarchitectonic and chemoarchitectonic characterization of the prefrontal cortical areas in the mouse. *Brain Struct. Funct.* **214,** 339–353 (2010).
34. Zong, H., Espinosa, J. S., Su, H. H., Muzumdar, M. D. & Luo, L. Mosaic analysis with double markers in mice. *Cell* **121,** 479–492 (2005).
35. Harris, K. D., Hirase, H., Leinekugel, X., Henze, D. A. & Buzsaki, G. Temporal interaction between single spikes and complex spike bursts in hippocampal pyramidal cells. *Neuron* **32,** 141–149 (2001).
36. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131,** 1–11 (2005).
37. Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *Information Theory*. *IEEE Trans. Inform. Theory* **49,** 1858–1860 (2003).
38. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482,** 85–88 (2012).
39. Brody, C. D. Correlations without synchrony. *Neural Comput.* **11,** 1537–1551 (1999).
40. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nature Rev. Neurosci.* **7,** 358–366 (2006).
41. Nikolić, D., Mureşan, R. C., Feng, W. & Singer, W. Scaled correlation analysis: a better way to compute a cross-correlogram. *Eur. J. Neurosci.* **35,** 742–762 (2012).
42. Marshall, L. *et al.* Hippocampal pyramidal cell-interneuron spike transmission is frequency dependent and responsible for place modulation of interneuron discharge. *J. Neurosci.* **22,** RC197 (2002).
43. Hangya, B., Li, Y., Muller, R. U. & Czurko, A. Complementary spatial firing in place cell-interneuron pairs. *J. Physiol. (Lond.)* **588,** 4165–4175 (2010).
44. Barthó, P. *et al.* Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *J. Neurophysiol.* **92,** 600–608 (2004).
45. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B* **63,** 411–423 (2001).
46. Kepecs, A., Uchida, N., Zariwala, H. & Mainen, Z. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455,** 227–231 (2008).

doi:10.1038/nature12171

Topographic diversity of fungal and bacterial communities in human skin

Keisha Findley[1], Julia Oh[1], Joy Yang[1], Sean Conlan[1], Clayton Deming[1], Jennifer A. Meyer[1], Deborah Schoenfeld[2], Effie Nomicos[2], Morgan Park[3], NIH Intramural Sequencing Center Comparative Sequencing Program†, Heidi H. Kong[2]* & Julia A. Segre[1]*

Traditional culture-based methods have incompletely defined the microbial landscape of common recalcitrant human fungal skin diseases, including athlete's foot and toenail infections. Skin protects humans from invasion by pathogenic microorganisms and provides a home for diverse commensal microbiota[1]. Bacterial genomic sequence data have generated novel hypotheses about species and community structures underlying human disorders[2–4]. However, microbial diversity is not limited to bacteria; microorganisms such as fungi also have major roles in microbial community stability, human health and disease[5]. Genomic methodologies to identify fungal species and communities have been limited compared with those that are available for bacteria[6]. Fungal evolution can be reconstructed with phylogenetic markers, including ribosomal RNA gene regions and other highly conserved genes[7]. Here we sequenced and analysed fungal communities of 14 skin sites in 10 healthy adults. Eleven core-body and arm sites were dominated by fungi of the genus *Malassezia*, with only species-level classifications revealing fungal-community composition differences between sites. By contrast, three foot sites—plantar heel, toenail and toe web—showed high fungal diversity. Concurrent analysis of bacterial and fungal communities demonstrated that physiologic attributes and topography of skin differentially shape these two microbial communities. These results provide a framework for future investigation of the contribution of interactions between pathogenic and commensal fungal and bacterial communities to the maintainenace of human health and to disease pathogenesis.

Since Hippocrates first described oral candidiasis in 400 BC, scientists have sought to explore the roles that commensal and pathogenic fungi and microbial communities have in human health and disease[8,9]. Culture-based studies have reported *Malassezia, Rhodotorula, Debaromyces, Cryptococcus* and, in some sites, *Candida* as fungal skin commensals[10]. Cutaneous fungal infections affect 29 million North Americans, but the role of dermatophytes in common toenail infections can be difficult to characterize using culture-based studies[11]. For other common skin disorders, such as seborrheic dermatitis (dandruff), fungal involvement remains incompletely understood[12,13]. Difficulty in establishing growth conditions[14,15] contribute to challenges to rapidly identify and direct treatment against pathogenic fungi.

To compare culture- and DNA-sequence-based identification of human skin-associated fungi, we obtained parallel samples from four skin sites of adult healthy volunteers (Supplementary Fig. 1 and Supplementary Table 1). We characterized isolates by morphological features and molecular markers. In total, we cultured more than 130 fungal isolates: 62 *Malassezia* (species *globosa, restricta* and *sympodialis*[13,16]), 25 *Penicillium* (species *chrysogenum* and *lanosum*) and 19 *Aspergillus* (species *candidus, terreus* and *versicolor*) (Supplementary Table 2). Five or fewer *Alternaria, Candida, Chaetomium, Chrysosporium, Cladosporium, Mucor, Rhodotorula* and *Trichophyton* isolates were cultured.

To explore fungal diversity with culture-independent methods, we prepared DNA from clinical swabs, and polymerase chain reaction (PCR)-amplified and sequenced two phylogenetic markers within the ribosomal RNA region: 18S rRNA and the intervening internal transcribed spacer 1 (ITS1) region[7,17,18]. We generated a custom ITS1 database based on sequences deposited in GenBank to classify sequences to genus level with greater than 97% accuracy (Supplementary Table 3). 18S rRNA sequences were classified using the SILVA database[19]. We determined the relative abundance of fungal genera of the occiput (back of head), nare (nostril), plantar heel and retroauricular crease (behind the ear). The genus *Malassezia* predominated in the retroauricular crease, nare and occiput; this was consistent across 18S rRNA and ITS1-characterized samples. Plantar heel was the most diverse site with representation of *Malassezia, Aspergillus, Cryptococcus, Rhodotorula, Epicoccum* and others (Supplementary Fig. 2). ITS1 sequencing enabled greater genus-level taxonomic resolution, reflecting the specificity of the genomic region and richness of the molecular database. Based on technical and analytic advantages, we selected the ITS1 region for subsequent sequencing and analyses of fungal diversity.

We generated more than 5 million ITS1 sequences from 10 healthy volunteers (from each of whom 14 skin-site samples were taken) (Supplementary Table 4). Both Ascomycetous and Basidiomycetous fungi were identified as normal skin flora. The genus *Malassezia* predominated at all 11 core-body and arm sites: antecubital fossa, back, external auditory canal, glabella, hypothenar palm, inguinal crease, manubrium, nare, occiput, retroauricular crease and volar forearm (Fig. 1). We explored *Malassezia* species-level resolution with a taxonomic data set that we developed based on reference ITS1 sequences and our human-associated *Malassezia* isolates. Pairwise comparisons of these *Malassezia* ITS1 sequences showed sequence identity within species to be greater than 91%, and identity between species to be 70 to 88%. These *Malassezia* sequences served as references within the phylogenetic pplacer[20] program to classify approximately 80 to 90% of *Malassezia* sequences per skin site to species level. Species-level identification revealed fungal specificity between body sites (Fig. 1). *M. restricta* predominated in external auditory canal, retroauricular crease and glabella, and *M. globosa* predominated on back, occiput and inguinal crease. Sites such as nares, antecubital fossa, volar forearm and hypothenar palm were characterized by multiple species (*M. restricta, M. globosa* and *M. sympodialis*). In total, we identified 11 of the 14 known *Malassezia* species among skin sites, suggesting that human skin is colonized with a wide range of *Malassezia*. Based on species-level resolution, we observed that fungal diversity is more dependent on body site than individual subject. ITS1 sequences also

[1]Genetics and Molecular Biology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Dermatology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [3]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland 20852, USA.

*These authors contributed equally to this work.

†Lists of participants and affiliations appear at the end of the paper.

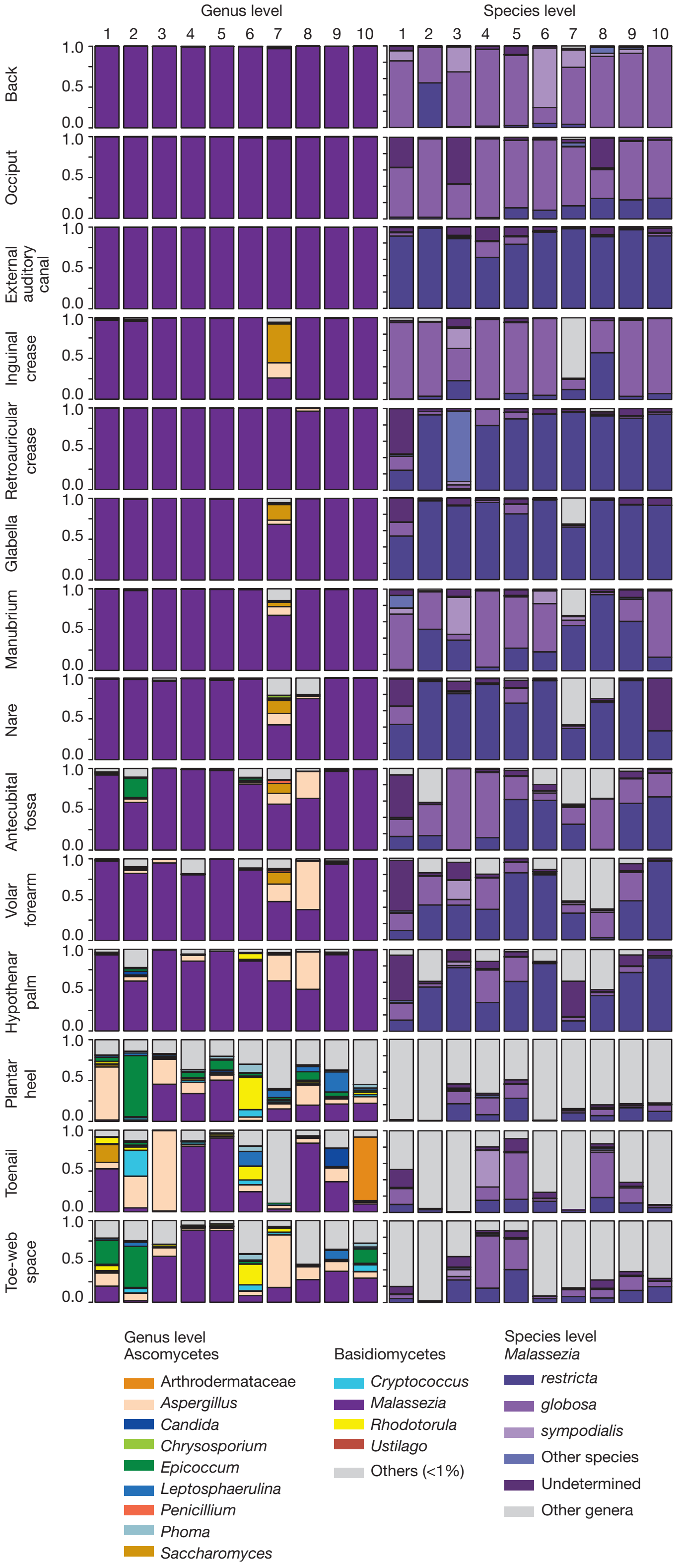


Figure 1 | Relative abundance of fungal genera and *Malassezia* species at different human skin sites. Fungal diversity of individual body sites of healthy volunteers (1–10) was taxonomically classified at the genus level, with further resolution of *Malassezia* species. For all body sites, the left side of the body was used, except for the right toenail of healthy volunteer 7.

matched *Candida* species *tropicalis*, *parapsilosis* and *orthopsilosis*, and *Cryptococcus* species *flavus*, *dimennae* and *diffluens*, which are considered to be part of the normal human flora and to be possible pathogens in wounds or immunocompromised patients[14].

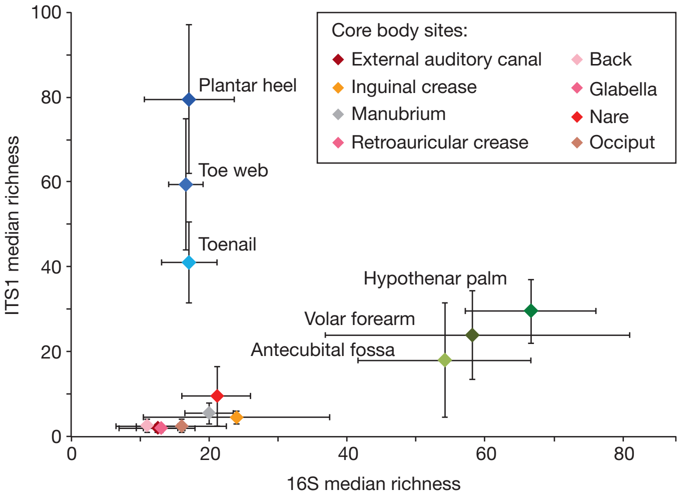


Figure 2 | Median richness of fungal and bacterial genera. Median taxonomic richness (or number of observed genera) of fungal and bacterial genera at 14 body sites. Error bars represent the median absolute deviation. The values for the core body sites retroauricular crease and glabella are identical and are therefore represented by a single data point on the graph and a shared colour in the key.

Substantially greater diversity was observed on three foot sites (plantar heel, toenail and toe web), in both the number of genera observed and the variation between individuals (Supplementary Fig. 3 and Supplementary Table 5). The fungal profile of one of the subjects (who we will refer to as healthy volunteer 7) was notably more diverse than other participants (Fig. 1). Healthy volunteer 7 completed a 2-month course of oral antifungal medication for a toenail infection 7 months before sampling. The remaining healthy volunteers reported no use of either oral or topical antifungal medication for at least 2 years before sampling. Healthy volunteer 7 is an outlier, but the additional genera that were identified (for example, *Aspergillus* and *Saccharomyces*) show that skin is capable of harbouring high fungal diversity. Although microbial sequencing is unable to determine causation, these data may suggest either that fungal community imbalance is associated with recurrent toenail infections or that alterations in fungal skin communities persist even 7 months after discontinuing antifungal medications. In comparison with culture-based analysis, ITS1 sequencing can provide a more complete view of the diversity of commensal microbiota, and also of potentially pathogenic microbiota.

To quantify and compare community similarity and taxonomic richness of skin sites, we assigned fungal sequences to taxonomic units based on genus-level phylogeny rather than percent sequence identity to obviate the high variation noted between species[21] with the latter metric. Plantar heel was the most complex fungal site (median richness of approximately 80 genera), and other foot sites had the next highest diversity (toe web and toenail, with approximately 60 and 40 genera, respectively; Fig. 2, Supplementary Table 6). Arm sites showed intermediate richness, ranging from 18 to 32 genera and core-body sites exhibited much lower richness, ranging from 2 to approximately 10 genera. Thus, regional location is a strong determinant of fungal richness. As observed in skin bacterial studies, left–right similarity within an individual was greater than between different individuals at the same body site (Supplementary Fig. 4 and Supplementary Table 7). To determine the temporal stability of the fungal microbiome, 6 healthy volunteers returned 1 to 3 months after initial sampling. Sites that showed *Malassezia* predominance at initial sampling displayed the same genus- and species-level predominance with strong community structure stability (Supplementary Figs 5 and 6, and Supplementary Table 8). Foot sites continued to show high diversity, perhaps reflecting frequent environmental exposure.

To explore the relationship between skin-associated fungi and bacteria, we sequenced the 16S rRNA gene from the same clinical samples. Consistent with previous studies[22,23], bacteria on healthy human skin

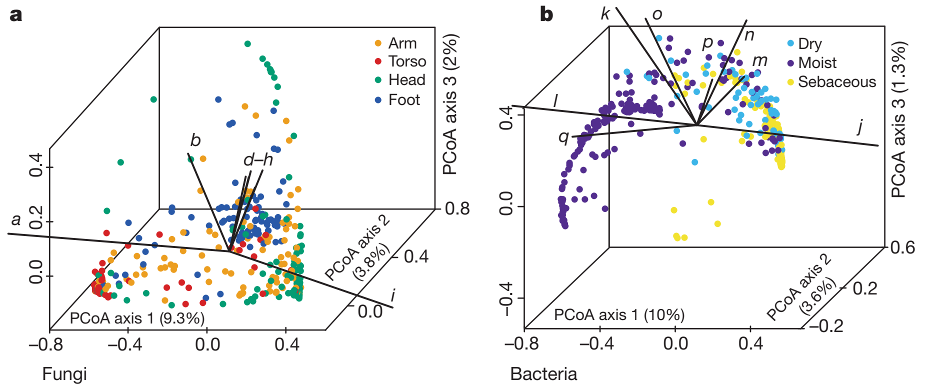


Figure 3 | Forces that shape fungal and bacterial communities. **a**, **b**, Principal coordinates analysis (PCoA) of the degree of fungal- and bacterial-community similarity at 14 body sites, based on predominant genera and species. Variation in fungal communities segregated strongly according to site location, with arm, foot, head and core-body sites forming discrete groups (**a**). Bacterial community structure was more dependent on site physiology (**b**). Axes that most significantly contribute to variation and their relative lengths are shown (these are defined in Supplementary Fig. 10 legend). For fungi, *M. restricta* (*i*; $\rho = 0.92$) and *M. globosa* (*a*; $\rho = -0.79$) are primary and opposing drivers of variation based on Spearman correlation with PCoA axis 1. For bacteria, *Propionibacterium* (*j*; $\rho = 0.95$) contributes to sebaceous site variation, whereas *Corynebacterium* (*l*; $\rho = -0.74$) and *Turicella* (*q*; $\rho = -0.56$) are the greatest contributors at moist sites based on Spearman correlation with PCoA axis 1.

were predominantly *Propionibacterium*, *Corynebacterium* and *Staphylococcus* (Supplementary Fig. 7). Similar to other moist skin sites, the toenail bacteria (not surveyed previously) were predominantly *Corynebacterium* and *Staphylococcus*. Interestingly, although healthy volunteer 7 was an outlier in terms of fungal diversity and membership, the bacterial profile was normal with respect to taxonomic and ecological measures of diversity (Supplementary Fig. 7). Directed studies may help elucidate how antibacterial and antifungal therapies perturb fungal and bacterial communities.

Bacterial and fungal richness were not linearly correlated, but were instead grouped into discrete clusters of sites from the same region; arm, foot and core-body (sites from the same regions had similar bacterial and fungal richness) (Fig. 2 and Supplementary Fig. 8). Arm sites displayed markedly greater bacterial diversity and lower fungal diversity than the foot and core-body sites. In contrast, foot sites displayed markedly greater fungal diversity with lower bacterial diversity than the arm and core-body sites. Core-body sites clustered together, and showed both lower bacterial and lower fungal diversity. These data reveal that the skin microbiome is complex and suggest that different characteristics shape skin bacterial and fungal communities.

Using principal coordinates analysis of community structure, we explored properties that may shape bacterial and fungal communities differentially. Consistent with previous studies[5], bacterial communities varied in the proportion of lipophilic bacteria (*Corynebacterium*, *Propionibacterium* and *Turicella*) and staphylococcal species, and grouped based on skin physiology into sebaceous, moist and dry sites (Fig. 3 and Supplementary Fig. 9). In contrast, fungal communities were segregated more clearly by site location than physiology, with foot, arm, head and torso sites forming discrete groups. Different *Malassezia* species drive variation in arms, torso and head, whereas a wide range of fungal genera drive variation in feet (Fig. 3 and Supplementary Fig. 9). Co-occurrence analysis of foot sites, based on Spearman correlation of fungal and bacterial taxonomic relative abundances (Supplementary Fig. 10), provided a preliminary evaluation of major fungal–bacterial associations. For example, a group of primarily Actinobacteria was anti-correlated with resident Ascomycota and Basidiomycota in contrast to a group of primarily Firmicutes and Proteobacteria that was positively correlated with these fungal taxa.

We observed that 20% (12 out of 60) of our study participants had clinical involvement (plantar-heel scaling, toe web scaling or toenail changes), consistent with possible fungal infections (Supplementary Table 1). Of the subjects with observed clinical involvement, positive mycological cultures were obtained from toenails (two samples) and plantar heel (one sample) (*Trichophyton*, *Penicillium* and *Aspergillus*). These observations are similar to the results of larger studies, which report signs of clinical involvement in up to 60% of feet of healthy individuals, and find 2 to 25% of cultures to be positive for fungi. The wide variation in prevalence of clinical involvement and positive fungal cultures is dependent on several factors, including population and climate[24,25]. As an initial inquiry into the aetiology of foot fungal disorders, we examined how observed clinical status (involved or uninvolved) at plantar heel, toe web or toenail affected fungal community structure. For uninvolved sites, interpersonal variation in community structure was highly consistent across all foot sites. In contrast, for sites with observed clinical involvement, similarity of community structure was much higher for plantar heel but much lower for toenail (Fig. 4). These data may suggest that there is a common fungal community shared among individuals with plantar-heel involvement, and high fungal diversity underlying toenail infections, but further studies are needed. These data sets can now be used to inform future clinical studies that examine microbial community shifts associated with fungal infections.

This systematic study clearly demonstrates that human skin surfaces are complex ecosystems, providing diverse environments for microorganisms that inhabit our bodies. Different factors determine bacterial and fungal communities, depending on the physiological properties of the skin. *Malassezia* species predominate on all core-body and arm sites. In contrast, foot sites display tremendous fungal diversity and markedly lower stability over time. Microbial community instability may provide an opportunity for potentially pathogenic microbes to establish disease. Plantar heels, toe webs and toenails are common sites of

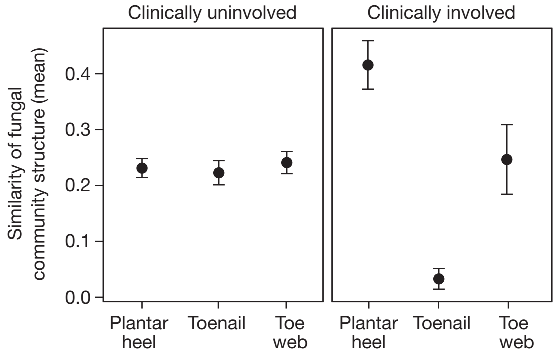


Figure 4 | Clinical involvement alters shared fungal community structure. Community structure measures the type and relative abundance of each genus. A value of 1 implies identical community structure and 0 implies dissimilar structures. Among uninvolved foot sites, community structure is fairly consistent at plantar heel, toenail and toe web sites. For involved sites, plantar heel has much greater shared community structure and toenails have much lower shared community structure. Error bars represent the s.e.m.

recurrent human fungal disease, which can be recalcitrant to treatment. This study also investigated fungal diversity at sites of predilection for other skin disorders, including seborrheic dermatitis, tinea cruris and subtypes of atopic dermatitis. With genomic advances, such as shotgun metagenomic sequencing, it is possible to begin to address interactions between microbes (bacterial–fungal, bacterial–bacterial, fungal–fungal) residing in these complex environments. The role of fungal commensals in educating the human immune system is gaining new appreciation in intestinal disease[26]. Further studies of healthy skin and dermatologic disorders are needed to explore these host–microbe interactions. In addition, antifungal medications, including azoles, echinocandins and amphotericin B, have potentially serious side effects such as liver or kidney damage[27]. Therefore, new treatment approaches are required to strategically target microbial dysbiosis and to combat the increasingly observed resistance against our current arsenal of antimicrobial therapies.

METHODS SUMMARY

Subject recruitment and sample preparation. This study was approved by the Institutional Review Board of the National Human Genome Research Institute (http://www.clinicaltrials.gov/ct2/show/NCT00605878) and all subjects provided written informed consent before participation. For fungal culturing studies, skin was scraped with a surgical blade and placed directly in media. For sequencing studies, swabs from skin and environmental negative controls were placed in MasterPure Yeast DNA purification kit (Epicentre) lysis buffer augmented with lysozyme. Proteinase K (Invitrogen) was added to pre-digest toenail clippings and incubated overnight with shaking at 55 °C. Steel beads (5 mm in diameter) were added to mechanically disrupt fungal cell walls using Tissuelyser (Qiagen) for 2 min at 30 Hz and then using the PureLink Genomic DNA Kit (Invitrogen). For 18S rRNA sequencing, each DNA was amplified with SR6 (5′-TACCTGG TTGATTCTGC) and SR1R (5′-TGTTACGACTTTTACTT) primers. For ITS1 sequencing, each DNA was amplified with adaptor plus 18SF (5′-GTAA AAGTCGTAACAAGGTTTC) and 5.8S1R plus barcode primers (5′-GTTCA AAGAYTCGATGATTCAC). For 16S rRNA sequencing, each DNA was amplified with adaptor plus V1_27F (5′-AGAGTTTGATCCTGGCTCAG) and V3_534R plus barcode primers (5′-CAGCACGCATTACCGCGGCTGCTGG).

Sequence classification and analyses. Sequences were pre-processed to remove primers and barcodes. Possible chimaeras were identified with UCHIME in mothur[21,28]. ITS1 sequences were classified to genus level with BLAST (basic local-alignment search tool) and the *k*-nearest neighbour algorithm in mothur. 18S rRNA sequences were classified using the SILVA v108 database[19]. 16S rRNA sequences were classified to genus level using the RDP classifier and training set 6[29]. We curated and aligned a reference library of *Malassezia* ITS1 type-strain sequences retrieved from GenBank augmented with those from our human-associated fungal cultures. This curated library was used as a reference to phylogenetically place and classify ITS1 sequences to species level within pplacer[20]. Sequence placement on the reference tree was visualized in Archaeopteryx using the 'guppy' command for classifications with a likelihood score of greater than or equal to 0.65. Full methods are found in the Supplementary Information.

Full Methods and any associated references are available in the online version of the paper.

Received 5 December 2012; accepted 9 April 2013.

Published online 22 May 2013.

1. Marples, M. (ed.) *The Ecology of the Human Skin* (Bannerstone House, 1965).
2. Grice, E. A. & Segre, J. A. The human microbiome: our second genome. *Annu. Rev. Genomics Hum. Genet.* **13,** 151–170 (2012).
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
4. Pflughoeft, K. J. & Versalovic, J. Human microbiome in health and disease. *Annu. Rev. Pathol.* **7,** 99–122 (2012).
5. Peleg, A. Y., Hogan, D. A. & Mylonakis, E. Medically important bacterial–fungal interactions. *Nature Rev. Microbiol.* **8,** 340–349 (2010).
6. Dollive, S. *et al.* A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol.* **13,** R60 (2012).
7. James, T. Y. *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443,** 818–822 (2006).
8. Ghannoum, M. A. *et al.* Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6,** e1000713 (2010).
9. Paulino, L. C., Tseng, C. H., Strober, B. E. & Blaser, M. J. Molecular analysis of fungal microbiota in samples from healthy human skin and psoriatic lesions. *J. Clin. Microbiol.* **44,** 2933–2941 (2006).
10. Roth, R. R. & James, W. D. Microbial ecology of the skin. *Annu. Rev. Microbiol.* **42,** 441–464 (1988).
11. Bickers, D. R. *et al.* The burden of skin diseases: 2004 a joint project of the American Academy of Dermatology Association and the Society for Investigative Dermatology. *J. Am. Acad. Dermatol.* **55,** 490–500 (2006).
12. Gaitanis, G., Magiatis, P., Hantschke, M., Bassukas, I. D. & Velegraki, A. The *Malassezia* genus in skin and systemic diseases. *Clin. Microbiol. Rev.* **25,** 106–141 (2012).
13. Saunders, C. W., Scheynius, A. & Heitman, J. *Malassezia* fungi are specialized to live on skin and associated with dandruff, eczema, and other skin diseases. *PLoS Pathog.* **8,** e1002701 (2012).
14. Larone, D. H. *Medically Important Fungi: A guide to identification* (ASM Press, 2002).
15. St-Germain, G. & Summerbell, R. *Identifying Fungi: A Clinical Laboratory Handbook* (Star Publishing Company, 2011).
16. Gioti, A. *et al.* Genomic insights into the atopic eczema-associated skin commensal yeast *Malassezia sympodialis*. *mBio* **4,** e00572–12 (2013).
17. Bruns, T. D. *et al.* Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. *Mol. Phylogenet. Evol.* **1,** 231–241 (1992).
18. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl Acad. Sci. USA* **109,** 6241–6246 (2012).
19. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–D596 (2013).
20. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11,** 538 (2010).
21. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
22. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).
23. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science* **324,** 1190–1192 (2009).
24. Cohen, A. D., Wolak, A., Alkan, M., Shalev, R. & Vardy, D. A. Prevalence and risk factors for tinea pedis in Israeli soldiers. *Int. J. Dermatol.* **44,** 1002–1005 (2005).
25. Perea, S. *et al.* Prevalence and risk factors of tinea unguium and tinea pedis in the general population in Spain. *J. Clin. Microbiol.* **38,** 3226–3230 (2000).
26. Iliev, I. D. *et al.* Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. *Science* **336,** 1314–1317 (2012).
27. Perfect, J. R., Lindsay, M. H. & Drew, R. H. Adverse drug reactions to systemic antifungals. Prevention and management. *Drug Saf.* **7,** 323–363 (1992).
28. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
29. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Heitman, A. Amend, Y. Shea, M. Turner, I. Brownell and M. Udey for helpful discussions. We thank J. Fekecs for assistance with the figures. This work was supported by the US National Institutes of Health (NIH) NHGRI and NCI Intramural Research Programs, and in part by NIH grant no. 1K99AR059222 (to H.H.K.). Sequencing was funded by grants from the NIH (1UH2AR057504-01 and 4UH3AR057504-02).

Author Contributions K.F., H.H.K. and J.A.S. designed the outline of the study. D.S. and E.N. recruited human subjects and assisted H.H.K. in sample collection for the experiment. K.F. and J.Y. assembled and curated the fungal database. J.A.M. and C.D. prepared the clinical samples for sequencing. The members of the NIH Intramural Sequencing Center Comparative Sequencing program carried out sequencing. K.F., J.O., S.C. and M.P. analysed sequence data. K.F., H.H.K. and J.A.S. drafted the manuscript with specific contributions from J.O., J.Y. and S.C. All authors read and approved the final version of the manuscript.

Author Information Sequence data from this study have been submitted to GenBank/EMBL/DDBJ under accession numbers KC669797–KC675175, and the Sequence Read Archive, and can be accessed through BioProject identification no. 46333. Patient and sample metadata have been deposited in the controlled-access Database of Genotypes and Phenotypes (dbGaP) under study accession phs000266.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.H.K. (konghe@mail.nih.gov) or J.A.S. (jsegre@nhgri.nih.gov).

NIH Intramural Sequencing Center Comparative Sequencing Program

Jesse Becker[1], Betty Benjamin[1], Robert Blakesley[1], Gerry Bouffard[1], Shelise Brooks[1], Holly Coleman[1], Mila Dekhtyar[1], Michael Gregory[1], Xiaobin Guan[1], Jyoti Gupta[1], Joel Han[1], April Hargrove[1], Shi-ling Ho[1], Taccara Johnson[1], Richelle Legaspi[1], Sean Lovett[1], Quino Maduro[1], Cathy Masiello[1], Baishali Maskeri[1], Jenny McDowell[1], Casandra Montemayor[1], James Mullikin[1], Morgan Park[1], Nancy Riebow[1], Karen Schandler[1], Brian Schmidt[1], Christina Sison[1], Mal Stantripop[1], James Thomas[1], Pam Thomas[1], Meg Vemulapalli[1] & Alice Young[1]

[1]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland 20852, USA.

METHODS

Subject recruitment and sampling. Healthy adult male and female volunteers of 18 to 40 years of age were recruited from the Washington, DC metropolitan region, United States, from September 2009 to September 2011. This natural history study was approved by the Institutional Review Board of the National Human Genome Research Institute (http://www.clinicaltrials.gov/ct2/show/NCT00605878) and all subjects provided written informed consent before participation. Subjects provided medical and medication history and underwent a physical examination. Exclusion criteria included history of chronic medical conditions, including chronic dermatologic diseases, and use of antimicrobial medication (antibiotic or antifungal treatments) 6 months before sampling (see Supplementary Table 1 for information on healthy volunteers). Bathing or showering with only non-antibacterial soap or cleansers was allowed during the 7 days before sample collection. No bathing, shampooing or moisturizing was permitted for 24 h before sample collection. Some healthy volunteers returned 1 to 3 months after their initial visit for follow-up sampling.

Fourteen skin sites representing a range of physiological characteristics and sites of predilection for fungus-associated dermatologic diseases were selected. Proximal and core-body sites were as follows: middle upper back, external auditory canal (inside the ear), retroauricular crease (behind the ear), occiput (back of scalp), glabella (central forehead, between eyebrows), inguinal crease (skin fold midway between hip and groin area), manubrium (upper central chest) and nare (inside the nostril). Distal body sites were as follows: antecubital fossa (inner elbow), volar forearm (mid-inner forearm), hypothenar palm (palm of hand, area closest to little finger), plantar heel (bottom of heel), toenail, and toe web (web-space between third and fourth toes) (Supplementary Fig. 1). All clinical findings observed at sampling sites were documented, including any scaling on the feet and toenail thickening, discoloration or subungal debris. Body sites with left–right symmetry (10 of the 14 body sites) were sampled bilaterally to calculate intrapersonal variation (see Supplementary Fig. 1 for sites sampled).

Fungal culturing and characterization. For fungal cultures, superficial skin scrapes were collected from a 4-cm^2 area with a sterile surgical blade and placed directly in media. Skin scrapings were spread on fungal culturing plates (under a laminar flow hood to minimize contamination) to isolate pathogenic and non-pathogenic fungi, including fastidious yeasts. Selective media containing antibiotic treatments to selectively suppress bacterial growth included: inhibitory mould agar with gentamicin (R01506); BHI agar with sheep blood, chloramphenicol and gentamicin (R01144); and Sabouraud dextrose agar, Emmons with chloramphenicol and cycloheximide (R01771) (Thermo Scientific) augmented with olive oil to promote *Malassezia* growth. Plates were incubated at 30 °C, checked daily for the first week and 2 to 3 days thereafter. Isolates that flourished in culture were re-streaked for single colonies, then subcultured to ensure purity and characterized by morphological features and molecular markers. DNA was extracted using the MasterPure Yeast DNA Purification Kit (Epicentre) according to the manufacturer's instructions with the addition of 5-mm steel beads to disrupt fungal cell wells mechanically (Qiagen). ITS1 and ITS2 regions were amplified from purified genomic fungal DNA using primers 18S-F (5′-GTAAAAGTCGTAACAAGGTTTC-3′) and 5.8S-1R (5′-GTTCAAAGAYTCGATGATTCAC-3′) for ITS1 and 5.8S-F (5′-GTGAATCATCGARTCTTTGAAC-3′) and 28S1-R (5′-TATGCTTAAGTTCAGCGGGTA-3′) for ITS2. PCR products were purified and sent to ACGT Inc. for sequencing and BLAST was carried out on the resulting amplicon sequence to identity each isolate[30].

Clinical sample collection, DNA extraction, PCR amplification and sequencing of 18S rRNA gene and ITS1. For DNA analyses, samples were collected, including negative controls, as described previously[31]. Catch-All Sample Collection Swabs (Epicentre) were used for skin-swab sample collection across all sites with the exception of the toenail (toenail clippings were collected)[32], and swabs were stored in lysis solution provided with the MasterPure Yeast DNA Purification Kit (Epicentre). To pre-digest the toenail clippings, Proteinase K (Invitrogen) was added to the sample and incubated overnight with shaking at 55 °C. Skin samples were incubated in yeast lysis buffer and lysozyme (20 mg ml^{-1}) for 1 h with shaking at 37 °C. Then, 5-mm steel beads were added to mechanically disrupt fungal cell walls using a Tissuelyser (Qiagen) for 2 min at 30 Hz. The Invitrogen PureLink Genomic DNA Kit (Invitrogen) was used for all subsequent steps.

For 18S rRNA amplicon sequencing, each DNA was amplified with SR6 (5′-TACCTGGTTGATTCTGC-3′) and SR1R (5′-TGTTACGACTTTTACTT-3′) primers. The following PCR conditions were used: 2.5 µl 10× AccuPrime Buffer II, 0.2 µl Accuprime Taq, 0.5 µl primer SR6 (20 µM), 0.5 µl primer SR1R (20 µM), and 4 µl of isolated microbial genomic DNA. PCR was carried out in duplicate when possible and a portion of the reaction was run on an agarose gel to verify the presence of the 18S PCR product. Cycle number was determined such that amplification was still in the linear range of the reaction and produced sufficient PCR product for cloning (maximum of 32 cycles). Negative controls on both the mock swab and water only (no template DNA) were performed with each set of amplifications to monitor procedures and reagents, respectively. The PCR product was ligated into the PCR4 TOPO vector (Invitrogen) according to the manufacturer's protocol. Of the resulting bacterial colonies, 384 per ligation were picked, plasmid DNA purified and inserts sequenced at NISC on an ABI 3730xl sequencer (Applied Biosystems) using M13 primers flanking the insert.

For ITS1 amplicon sequencing, each DNA was amplified with adaptor plus 18SF (5′-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGTAAAAGTCGTAACAAGGTTTC) and 5.8S-1R plus barcode (5′-GTTCAAAGAYTCGATGATTCAC) primers[33]. The following PCR conditions were used: 2.5 µl 10× AccuPrime Buffer II, 0.2 µl Accuprime Taq (Invitrogen), 0.1 µl primer B adaptor plus 18SF (100 µM), 2 µl primer 5.8S-1R plus barcode (5 µM), and 4 µl of isolated microbial genomic DNA. The PCR was carried out in duplicate for 32 cycles. Duplicate amplicons were combined, purified (Agencourt AMPure XP-PCR Purification Kit; Beckman Coulter), and quantified (QuantIT dsDNA High-Sensitivity Assay Kit; Invitrogen). An average of approximately 8 ng DNA of 94 amplicons were pooled together, purified (MinElute PCR Purification Ki; Qiagen) and sequenced on a Roche 454 GS20/FLX platform with Titanium chemistry (Roche). Flowgrams were processed with the 454 Basecalling pipeline (v.2.5.3).

For 16S rRNA amplicon sequencing, each DNA was amplified with adaptor plus V1_27F (5′-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGAGAGTTTGATCCTGGCTCAG) and V3_534R plus barcode primers (5′-CAGCACGCATTACCGCGGCTGCTGG)[33]. The following PCR conditions were used: 2 µl 10× AccuPrime Buffer II, 0.15 µl Accuprime Taq (Invitrogen), 0.04 µl adaptor plus V1_27F (100 µM), 2 µl primer V3_354R plus barcode (2 µM), and 2 µl of isolated microbial genomic DNA. PCR was carried out in duplicate for 30 cycles and then cleaning up of PCR, amplicon pooling of approximately 10 ng DNA, purification, and sequencing were performed as described above for ITS1.

Custom-generated fungal ITS1 reference database. ITS sequences were extracted from GenBank using the query: (ITS1[All Fields] OR ITS2[All Fields] OR 5.8S[All Fields]) AND Fungi[All Fields] NOT 'uncultured'[All Fields]. Taxonomy classifications associated with sequences were recorded as strings in the following order: kingdom, phylum, class, order, family and genus, and recorded as unclassified if the levels were not clearly defined. Sequence classification was curated manually and any discrepancies in taxonomy strings were resolved using the Taxonomy Database in Pubmed. When both anamorphic (asexual) and teleomorphic (sexual) names were represented for a species within GenBank, the strings were curated manually and the anamorphic taxonomic nomenclature was selected. The sequences were then clustered to 95% sequence identity using CD-HIT[34]. Representative sequences were chosen by CD-HIT and a consensus taxonomy string was generated for the sequence, starting from the highest level (kingdom) and moving to the lowest level (genus). If the most highly represented classification was twice as frequent as the next one, this classification was chosen for the level. If no classification satisfied this criterion, this and all lower levels were set as unclassified. Sequences that were clearly misclassified as fungi were removed from the curated database.

Sequence classification and analyses. Sequences were pre-processed to remove primers and barcodes. Possible chimaeras created during PCR amplification were identified with UCHIME in mothur[35,36]. Input category for 'reference' was set to self and included in the names file to check for chimaeras, thereby using more abundant sequences as references[36]. With the ITS database described above as the reference, these chimaera-checked sequences were classified to the genus level with the BLAST option and the *k*-nearest neighbour algorithm in mothur[36]. 16S rRNA sequences were classified to the genus or species level using the RDP classifier with training set (v.6) as described previously[37]. Sequences were assigned to taxonomic units based on their genus-level phylogenetic classification. R statistical software was implemented to generate plots representing the relative abundance of fungal genera.

Community richness (Chao1), diversity (Shannon Index), membership (Jaccard Index) and structure (Theta Index) were calculated within mothur as described previously after using a subsampling cut-off of 1,000 sequences per sample[31,38,39]. Diversity indices for left and right symmetric sites were averaged for body sites with bilateral symmetry. All statistical analyses are represented as the standard error of the mean unless otherwise indicated.

18S rRNA sequence classification. To pre-process the 18S rRNA sequences, traces were base-called using Phred (v.0.990722.g), trimmed with Crossmatch, and each clone assembled using Phrap (v.0.990329). The default parameters were used except that the force level was 9 and the mismatch penalty was −1 (refs 40, 41). For approximately 15% of read pairs, the overlap was not sufficient for *de novo* assembly and a scaffolded assembly was attempted. Scaffolded assembly was carried out using the AmosCmp16Spipeline (available from http://microbiomeutil.sourceforge.net) and non-redundant reference sequences from the SILVA small subunit rRNA database[42]. 18S rRNA sequences were classified using the SILVA v.108 database. R was implemented to generate plots representing the relative abundance of fungal genera.

***Malassezia* sequence classification to the species level.** To classify *Malassezia* ITS1 sequences from the genus to species level, we used the get.lineage command in mothur to retrieve only *Malassezia* sequences. As an internal check, these skin-associated *Malassezia* ITS1 sequences were aligned with the *Malassezia* reference package in mothur, and all discrepancies were resolved. We next curated and aligned a reference library of *Malassezia* type-strain ITS1 sequences retrieved from GenBank and augmented by those from the fungal cultures described above. Two to ten representatives were included in the database for *Malassezia* species (*M. globosa, M. restricta, M. sympodialis, M. slooffiae, M. furfur, M. pachydermatis, M. dermatis, M. yamatoensis, M. obtusa, M. japonica* and *M. nana*) for a total of 52 ITS1 sequences. Sequences were aligned with MUSCLE to generate the reference alignment[43].

This curated library was used as the reference to place and classify novel skin-associated *Malassezia* ITS1 sequences phylogenetically to the species level with the software package pplacer[44]. Sequence placement on the reference tree along with confidence scores were visualized in Archaeopteryx using the 'guppy' command[44,45]. The 'guppy classify' output and a lightweight SQL database were used to make and store taxonomic classifications. A likelihood score of at least 0.65 was used for classifications produced by the guppy classify command. Finally, classifications were converted into mothur-compatible taxonomic strings to create the tax.summary file for community-based analyses as above. Similarly, species-level designations for bacterial *Staphylococcus* sequences were generated using pplacer with a 16S rRNA reference database built from rRNA records extracted from RefSeq genomes (as of April 2012) and RDP type species sequences (Release 10, Update 24)[37].

ITS1 and 16S rRNA comparisons. Taxonomic units were defined from genus- and, where available, species-level ITS1 and 16S rRNA phylotypes. Groups were each subsampled to 1,800 sequences and the Yue–Clayton theta index generated to compare the similarity between communities. Principal coordinate analysis of the theta index was performed and the Spearman correlation of the relative abundance of each taxonomic unit versus the top three axes was calculated to assess how each taxonomic unit contributed to variation along the axes.

Co-occurrence of bacteria and fungi was assessed by calculating the partial Spearman correlation of the relative abundances of the different taxa, adjusted for multiple within-patient measurements. Calculations were performed on Fisher-transformed *r* values. Comparisons were limited to those taxa that occurred in more than 25% of samples for either ITS1 or 16S rRNA, and for ITS1, if mean abundance across all samples exceeded 0.25%. Owing to the relatively high fungal diversity found at foot sites, only foot sites (plantar heel, toenail and toe web) were used.

30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
31. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science* **324,** 1190–1192 (2009).
32. Grice, E. A. *et al.* A diversity profile of the human skin microbiota. *Genome Res.* **18,** 1043–1050 (2008).
33. Lennon, N. J. *et al.* A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol.* **11,** R15 (2010).
34. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
35. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
36. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
37. Conlan, S., Kong, H. H. & Segre, J. A. Species-level analysis of DNA sequence data from the NIH Human Microbiome Project. *PLoS ONE* **7,** e47075 (2012).
38. Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res.* **22,** 850–859 (2012).
39. Yue, J. C. & Clayton, M. K. A similarity measure based on species proportions. *Comm. Statist. Theory Methods* **34,** 2123–2131 (2005).
40. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
41. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
42. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
43. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
44. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11,** 538 (2010).
45. Han, M. V. & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10,** 356 (2009).

LETTER

doi:10.1038/nature12175

Neutrophil swarms require LTB4 and integrins at sites of cell death *in vivo*

Tim Lämmermann[1], Philippe V. Afonso[2], Bastian R. Angermann[1], Ji Ming Wang[3], Wolfgang Kastenmüller[1,4], Carole A. Parent[2] & Ronald N. Germain[1]

Neutrophil recruitment from blood to extravascular sites of sterile or infectious tissue damage is a hallmark of early innate immune responses, and the molecular events leading to cell exit from the bloodstream have been well defined[1,2]. Once outside the vessel, individual neutrophils often show extremely coordinated chemotaxis and cluster formation reminiscent of the swarming behaviour of insects[3–11]. The molecular players that direct this response at the single-cell and population levels within the complexity of an inflamed tissue are unknown. Using two-photon intravital microscopy in mouse models of sterile injury and infection, we show a critical role for intercellular signal relay among neutrophils mediated by the lipid leukotriene B4, which acutely amplifies local cell death signals to enhance the radius of highly directed interstitial neutrophil recruitment. Integrin receptors are dispensable for long-distance migration[12], but have a previously unappreciated role in maintaining dense cellular clusters when congregating neutrophils rearrange the collagenous fibre network of the dermis to form a collagen-free zone at the wound centre. In this newly formed environment, integrins, in concert with neutrophil-derived leukotriene B4 and other chemoattractants, promote local neutrophil interaction while forming a tight wound seal. This wound seal has borders that cease to grow in kinetic concert with late recruitment of monocytes and macrophages at the edge of the displaced collagen fibres. Together, these data provide an initial molecular map of the factors that contribute to neutrophil swarming in the extravascular space of a damaged tissue. They reveal how local events are propagated over large-range distances, and how auto-signalling produces coordinated, self-organized neutrophil-swarming behaviour that isolates the wound or infectious site from surrounding viable tissue.

Neutrophil swarming has been observed using intravital microscopy in inflamed, infected or sterilely wounded tissues[3–11,13], and a series of sequential phases have been described[3,4]: (1) initial chemotaxis of individual neutrophils close to the damage, followed by (2) amplified chemotaxis of neutrophils from more distant interstitial regions, leading to (3) neutrophil clustering. To study the molecules controlling these distinct neutrophil-response phases, we used an inducible model of sterile skin injury in which a brief intense two-photon laser pulse causes focal, dermis-restricted tissue damage (Supplementary Fig. 2)[4]. We were specifically interested in how neutrophils coordinate swarming in the extravascular space, so we performed two-photon intravital microscopy (2P-IVM) of neutrophils that had already exited blood vessels and entered a mildly inflamed dermis before laser damage. Focal injury induced substantial interstitial chemotaxis of lysozyme 2–green fluorescent protein (Lyz2–GFP, also known as LysM–GFP)-positive neutrophils/monocytes that lasted ~25–40 min before cells accumulated in a cluster at the damage site and recruitment stopped (Fig. 1a). The dynamic behaviour of neutrophils differed from CX3CR1-positive macrophages/monocytes in the same environment, with neutrophils immediately showing highly directed chemotaxis towards the wound centre at high speeds (10–20 μm min^{-1}) and the CX3CR1-positive cells migrating at slower speeds (3–5 μm min^{-1}) and undergoing

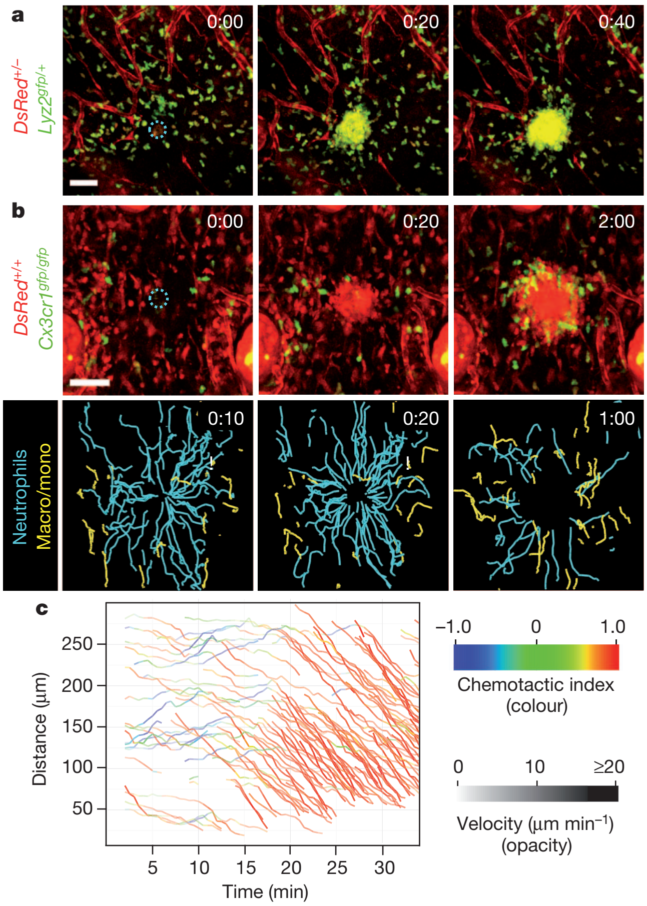


Figure 1 | Neutrophil extravascular swarming dynamics. 2P-IVM on intact ear dermis of anaesthetized mice. Interstitial cell recruitment towards focal damage (blue dotted circle) was recorded. **a**, **b**, Time-lapse sequence of endogenous innate immune cell dynamics in *DsRed*$^{+/-}$ *Lyz2*$^{gfp/+}$ *Tyr*$^{c-2J/c-2J}$ mice (myelomonocytic cells in green–yellow, stroma in red) (**a**) and *DsRed*$^{+/+}$ *Cx3cr1*$^{gfp/gfp}$ *Tyr*$^{c-2J/c-2J}$ mice (macrophages/monocytes in green, neutrophils and stroma in red) (**b**, top). Cell tracks over the last 10 min ($n = 4$) (**b**, bottom). Scale bars, 50 μm. Time, h:min. **c**, Distance–time plot (DTP) of intradermal (i.d.) injected bone marrow neutrophils; individual cell-migration paths towards the damage site are each highlighted with instantaneous chemotactic index (colour) and velocity (opacity). Representative experiment of $n = 169$.

[1] Laboratory of Systems Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892-0421, USA. [2] Laboratory of Cellular and Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4256, USA. [3] Laboratory of Molecular Immunoregulation, Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland 21702-1201, USA. [4] Institutes of Molecular Medicine and Experimental Immunology (IMMEI), University of Bonn, 53105 Bonn, Germany.

a chemotactic response only after an initial neutrophil cluster had formed. In contrast to neutrophils, macrophages/monocytes never entered the developing cell cluster, but assembled around it during cessation of the neutrophil swarm growth (Fig. 1b, Supplementary Fig. 3a and Supplementary Video 1). As reported previously, the early dynamics of endogenous neutrophils were biphasic (Supplementary Fig. 3b) and similar for neutrophils that were isolated from mouse bone marrow (Fig. 1c and Supplementary Video 2) or human peripheral blood (Supplementary Fig. 3c) and then injected into the dermis. In 79% of the experiments (169 out of 213), a first phase (1–15 min) of initial neutrophil chemotaxis close to the damage was followed by a dramatic second phase of substantial neutrophil recruitment from distant sites involving markedly increased directionality and speed (Fig. 1c). We noticed the appearance of cell fragments around developing neutrophil clusters, suggesting that cell death could lead to release of components driving neutrophil swarming[14]. Using propidium iodide, we detected a clear kinetic correlation between the death of a few neutrophils at the damage site and the onset of the amplified second phase of neutrophil recruitment (Fig. 2a, Supplementary Figs 4 and 5a and Supplementary Video 3). Most of the neutrophils in the developing cell cluster remained viable (Supplementary Fig. 5b), suggesting that the death of only a small number of neutrophils was sufficient to drive the substantial chemotactic response of the neutrophil population.

Although neutrophil death appeared to serve as a catalyst for swarming, it was unclear how such a signal would propagate through the structurally dense tissue in just a few minutes to initiate acute chemotaxis of neutrophils at sites more than 300 μm away from the core lesion, and also maintain directional migration for ~25–40 min. To investigate the nature of this signal, we performed intradermal co-injection experiments with control and gene-deficient neutrophils and imaged them side by side in the same tissue volume in each experiment. The experimental set-up eliminated problematic quantitative comparisons based on measurements of wild-type and mutant cells imaged in different experiments that can be influenced by varying tissue composition in the imaging volume (Supplementary Fig. 6). More importantly, this protocol bypassed the extravasation step, which allowed for the study of neutrophils depleted of molecules essential for vessel exit.

Neutrophils express more than 30 cell-surface receptors for various attractants, which upon activation and signalling rearrange the cytoskeleton to yield a functionally polarized neutrophil that is poised for directed migration[15,16]. In these cells, most G-protein-coupled receptors (GPCRs) act through the predominantly expressed $G\alpha_i$ isoforms $G\alpha_{i2}$ and $G\alpha_{i3}$. We found that only neutrophils genetically lacking $G\alpha_{i2}$ ($Gnai2^{-/-}$), but not $G\alpha_{i3}$ ($Gnai3^{-/-}$), show impaired interstitial chemotaxis (Supplementary Fig. 7)[17]. As $G\alpha_{i2}$ couples to several neutrophil GPCRs, we next investigated the function of individual GPCRs. Most knockouts for single receptor genes tested (*Fpr1*, *Fpr2*, *Cxcr2*, Supplementary Fig. 8 and Supplementary Video 4; *C5ar1*, *Ccr1*, *Ccr2*, *Ccr5*, *Cxcr6*, *Ptafr*, *P2ry2*, data not shown)—many of which have been previously reported to have important roles in inflammatory, infectious or autoimmune conditions (Supplementary Information)—showed normal interstitial chemotaxis, indicating that their respective ligands are not uniquely involved in the recruitment phase of the swarming response. Similarly, migration of neutrophils lacking receptors that detect the presence of danger signals (*Myd88*, *Il1r1*, *Tnfrsf1a* and *Il1r1*, *P2rx7*) was normal (data not shown). However, neutrophils lacking the high-affinity receptor for leukotriene B4 (LTB4) showed a uniquely impaired swarming response. During early phases (<15 min), $Ltb4r1^{-/-}$ neutrophils close to the damage site still performed chemotaxis towards the wound, whereas cells from more distant sites were only poorly recruited (Fig. 2b, Supplementary Fig. 9a and Supplementary Video 5), a phenotype that became even more striking when

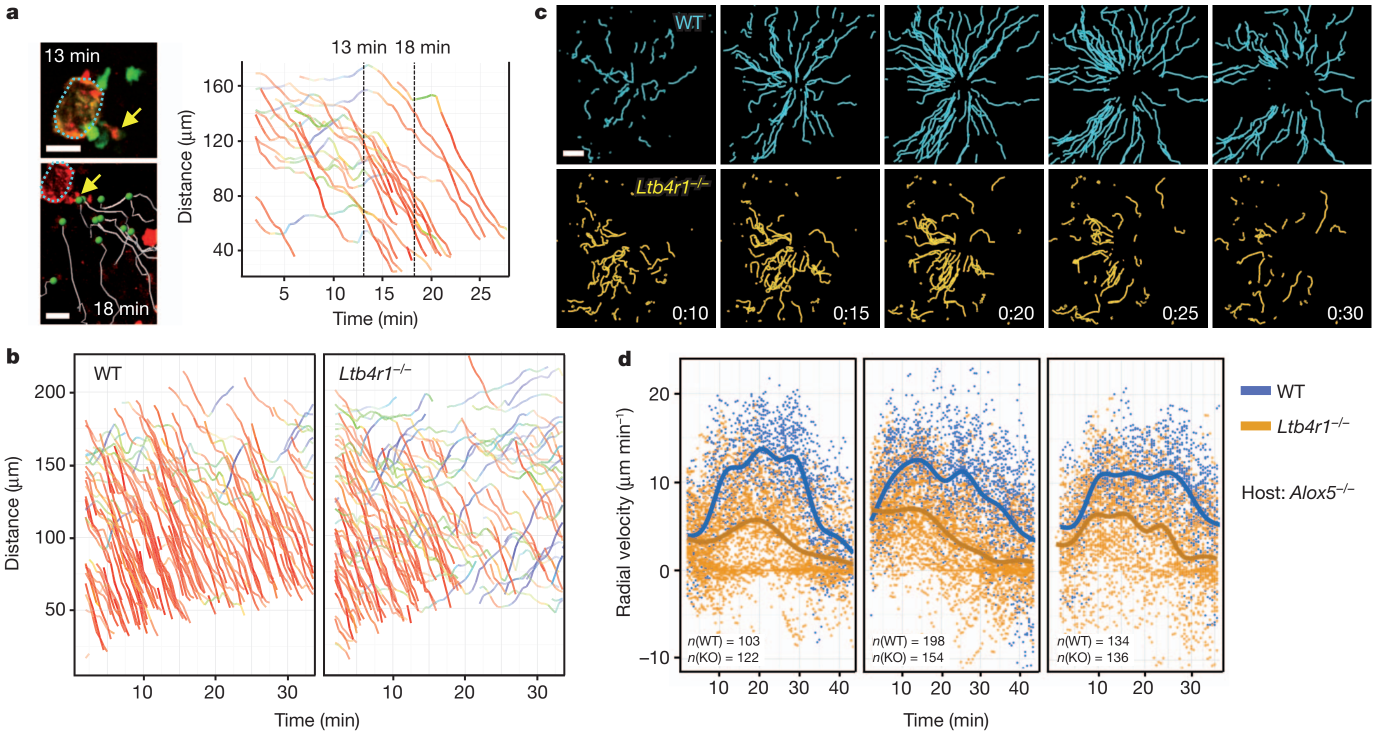


Figure 2 | LTB4 promotes neutrophil recruitment from distant sites. a, 2P-IVM images of a single neutrophil becoming propidium iodide-positive (arrow) at 13 min and its correlation to neutrophil-amplified chemotaxis (white tracks). DTP analysis for migration paths coloured as in Fig. 1. **b**, Comparative analysis of interstitial recruitment after i.d. co-injection of $Ltb4r1^{-/-}$ and wild-type (WT) neutrophils into $Tyr^{c\text{-}2J/c\text{-}2J}$ mice. DTP of one representative experiment ($n = 8$). **c**, Time-course of migration tracks towards 10-μm damage. Time, h:min. Scale bars, 20 μm (**a**), 50 μm (**c**). Track durations, 5 min (**a**), 10 min (**c**). **d**, Comparative analysis of interstitial recruitment after i.d. co-injection of $Ltb4r1^{-/-}$ knockout (KO) and wild-type neutrophils into $Alox5^{-/-}$ $Tyr^{c\text{-}2J/c\text{-}2J}$ mice. Radial velocity–time plots with regression lines showing the recruitment dynamics for three individual experiments (of $n = 7$). n (in graph) indicates the number of analysed tracks. Each dot represents the instantaneous radial velocity for one cell at that time point.

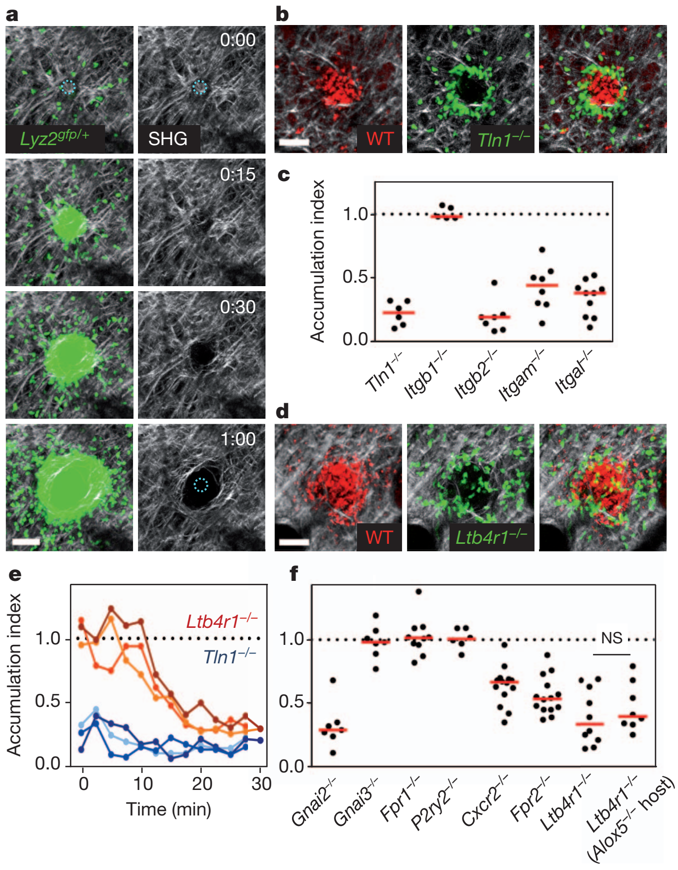


Figure 3 | Integrin and GPCR signalling at the neutrophil cluster. a, After focal damage (blue dotted circle), congregating neutrophils rearrange collagenous fibres (visualized by collagen second harmonic generation, SHG). Time, h:min. Scale bar, 50 μm. **b**–**f**, Comparative analysis of neutrophil clustering after i.d. co-injection of gene-deficient and wild-type neutrophils into *Tyr*$^{c\text{-}2J/c\text{-}2J}$ or *Alox5*$^{-/-}$ *Tyr*$^{c\text{-}2J/c\text{-}2J}$ mice. **b**, **d**, 2P-IVM images at the end point of the clustering response. Scale bars, 50 μm. **c**, **f**, Accumulation index as a quantitative parameter for neutrophil entry into the collagen-free wound centre. Each dot represents analysis of one damage site. Median in red. NS, non-significant (Mann–Whitney *U*-test). **e**, Time-course of neutrophil accumulation in the wound centre. Three representative experiments are presented for each gene deficiency.

the damage size was smaller (Fig. 2c, Supplementary Fig. 10 and Supplementary Video 6). This reduction in neutrophil recruitment was mainly due to impaired chemotaxis rather than chemokinesis, and was most obvious during the late phases of the swarming response (>15 min) (Supplementary Fig. 9b).

In inflamed tissues, LTB4 can originate from several cellular sources. Earlier studies implicated a role for neutrophil-derived LTB4 in the initial stages of neutrophil recruitment from blood into inflamed tissues and disease progression[18–22]. To examine whether neutrophils auto-amplified the swarming response through LTB4 production and signalling, we performed injection studies into 5-lipoxygenase-deficient mice (*Alox5*$^{-/-}$) that cannot synthesize leukotrienes, such that only injected neutrophils could act as an LTB4 source. When co-injected with control cells, *Ltb4r1*$^{-/-}$ neutrophils again showed impaired interstitial chemotaxis and recruitment from distant sites (Fig. 2d, Supplementary Fig. 11 and Supplementary Video 7). When LTB4-secreting wild-type neutrophils were injected alone, they migrated to the injury from distances >200 μm and formed large clusters. By contrast, when we injected *Alox5*$^{-/-}$ neutrophils alone into *Alox5*$^{-/-}$ hosts so that no leukotrienes were present, only neutrophils close to the damage site (<100 μm) were transiently recruited, resulting in small neutrophil clusters (Supplementary Fig. 12). Although these experiments establish an important role for neutrophil-derived LTB4 in recruiting neutrophils from distant sites in a feed-forward manner, they also show the existence of short-range chemotactic signals from the initial tissue injury and/or locally dying cells, as even in the absence of leukotrienes small clusters can form. We conclude that several of these initial primary factors not only act as short-range chemotactic signals, but also induce LTB4 secretion in neutrophils to acutely enhance the radius of neutrophil recruitment (Supplementary Fig. 1, panels 1–4), consistent with previous data on multiple LTB4-inducing factors[22–25]. This model is in agreement with earlier *in vitro* studies showing that primary signals induced secretion of LTB4 that acts as a signal relay molecule to improve chemotaxis of a whole neutrophil population[23].

Because LTB4 not only induces intracellular polarity to direct chemotaxis, but can also regulate neutrophil adhesiveness by activating integrin receptors[26,27], we next targeted integrin functionality by injecting neutrophils deficient for talin (*Tln1*$^{-/-}$). Talin interaction with integrin cytoplasmic domains is crucial for integrin activation, substrate binding and coupling of filamentous actin to adhesion sites[28]. Because interstitial chemotaxis of *Tln1*$^{-/-}$ cells was unimpaired (Supplementary Fig. 13 and Supplementary Video 8), high-affinity integrin function was dispensable for neutrophil interstitial migration *in vivo*, thus confirming earlier studies with dendritic cells in skin explants[12]. Together, our results demonstrate that LTB4 has a non-redundant role in directing tissue-migrating neutrophils to sites of tissue injury by activating chemotaxis signal pathways.

We next sought to analyse the subsequent step of swarming when neutrophils accumulate and form substantial cell clusters. Visualization of the dense dermal collagen network using the second harmonic generation signal showed clearance of visible fibres from the core of the wound where the densest clustering occurred (Fig. 3a), an effect that was not due to thermal damage (Supplementary Fig. 14a, b). Congregating neutrophils appeared to physically exclude collagen fibres from the core of the cell infiltrate (Supplementary Video 9), although some limited proteolysis of extracellular matrix structures cannot be ruled out. Notably, monocytes always aligned along collagen fibres at the outer edges of the cell cluster, suggesting that neutrophils are specifically adapted to enter the collagen-free centre (Supplementary Fig. 14c). When we visualized both endogenous neutrophils and the cellular actin cortex of injected Lifeact–GFP$^{+/-}$ neutrophils in the collagen-free wound core, we could clearly observe neutrophils migrating in contact with each other with motion paths and focal filamentous actin accumulations that were suggestive of adhesive interactions (Supplementary Video 10). Indeed, when *Tln1*$^{-/-}$ or *Itgb2*$^{-/-}$ (β2 integrin-deficient) neutrophils were co-injected with control cells, they showed a striking phenotype, as they were completely unable to enter the collagen-free zone and accumulated at the edge of the wild-type neutrophil cluster (Fig. 3b, c and Supplementary Videos 11, 12). Deficiency in either LFA-1 (*Itgal*$^{-/-}$) or Mac-1 (*Itgam*$^{-/-}$) had a measurable effect on central accumulation, suggesting that both were involved in cell adhesion and movement within the collagen-free zone (Fig. 3c and Supplementary Fig. 15a). These unexpected findings indicate that high-affinity integrins are critical for neutrophil accumulation in the collagen-free wound centre, which fundamentally differs in its extracellular architecture from the surrounding intact dermal interstitium where leukocyte migration does not require high-affinity integrin function.

We next examined the role of GPCR signals in regulating neutrophil aggregation at the wound. *Gnai2*$^{-/-}$ neutrophils were excluded from the central neutrophil cluster over time, whereas *Gnai3*$^{-/-}$ accumulated normally (Supplementary Fig. 15b and Supplementary Video 13). When investigating individual GPCRs, we detected impaired aggregation for neutrophils lacking CXCR2, FPR2 or LTB4R1, with the latter showing the strongest impact on the aggregation response (Fig. 3d, f and Supplementary Video 14; other tested GPCR or receptors had no effect, Supplementary Fig. 15). In contrast to adhesion-deficient *Tln1*$^{-/-}$ cells, *Ltb4r1*$^{-/-}$ neutrophils were intermixed with wild-type cells at the earliest aggregation stages. Over time, however, *Ltb4r1*$^{-/-}$ neutrophils were excluded from the growing cluster (Fig. 3e and Supplementary Fig. 16a). Similar results were obtained when *Ltb4r1*$^{-/-}$ and wild-type neutrophils were injected into *Alox5*$^{-/-}$ hosts (Fig. 3f and Supplementary Video 14), indicating that neutrophils secrete LTB4 to regulate clustering in a feed-forward manner. Detection of ligands for other cluster-mediating receptors, such as CXCL2 for CXCR2 and CRAMP for FPR2, in accumulating neutrophils (Supplementary Fig. 16b) suggests that signalling through multiple GPCR, especially LTB4R1, may act to increase neutrophil aggregation to maintain tight association and motility for continued uptake of cellular wound debris or potential pathogens in the cellular cluster (Supplementary Fig. 1, panels 5–8).

Our findings thus identify a crucial dual function for LTB4 in both the recruitment and the clustering phases of the swarm response to sterile injury. To test the relevance of these findings in an infectious situation, we investigated the role of LTB4 during transient neutrophil swarming in infected lymph nodes[3]. We have shown previously that *Pseudomonas aeruginosa* induces cell death of subcapsular macrophages, which subsequently leads to neutrophil recruitment to the lymph node[29]. Whereas endogenous neutrophils in control mice formed several large but transient cell clusters and were highly chemotactic between the competing transient clusters, *Ltb4r1*$^{-/-}$ neutrophils were slower and formed only very small, if any, clusters (Fig. 4a and Supplementary Video 15). Consequently, neutrophil cluster diameter and persistence were significantly lower in the absence of LTB4 signalling (Fig. 4b, c), thus confirming the results of the controlled inducible model of focal laser skin injury and extending them to other tissues, forms of cell death and infectious conditions.

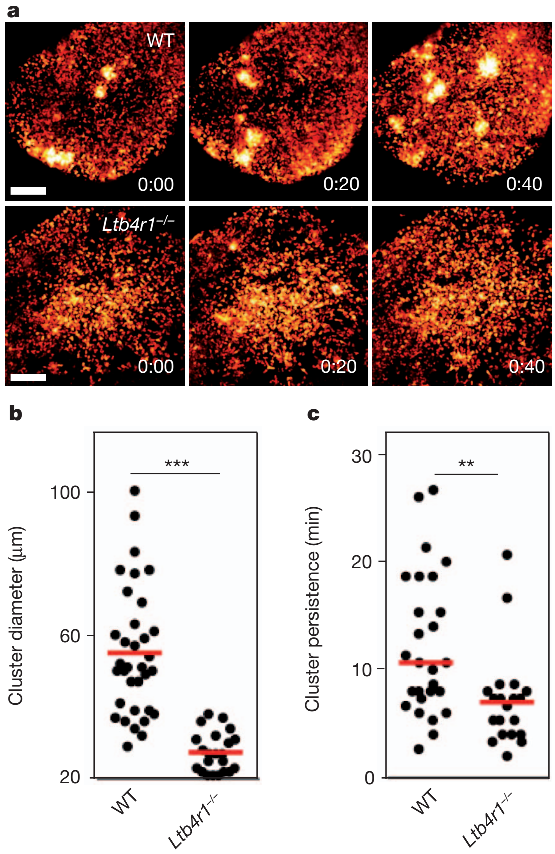


Figure 4 | LTB4 requirement for swarming in infected lymph nodes. Mice were infected with *P. aeruginosa*–GFP in the footpad and 2P-IVM was then performed on the draining popliteal lymph nodes at the times indicated. **a**, Time-lapse sequence of infected subcapsular sinuses of wild-type *Lyz2*$^{gfp/+}$ (top) and *Ltb4r1*$^{-/-}$ *Lyz2*$^{gfp/+}$ (bottom) mice. Analysis was performed when comparable neutrophil numbers were in the sinus (wild type, 3 h; *Ltb4r1*$^{-/-}$, 4.5 h). Neutrophil–GFP signal is pseudo-coloured (heat map) to indicate neutrophil clusters (white). Time, h:min. Scale bars, 100 μm. **b**, **c**, Quantification of cluster diameter (**b**) and persistence (**c**). Data pooled from three independent experiments. **b**, $***P < 0.001$ (Student's *t*-test); **c**, $**P < 0.01$ (Mann–Whitney *U*-test). Red bars, mean (**b**), median (**c**).

The chemoattractants that direct the neutrophil tissue response can stem from the injured tissue, resident cells, recruited blood-derived leukocytes and potential pathogens. The specific mixture of these various signals will determine the neutrophil chemotactic response at each inflammatory or infectious site. In large sterile liver injuries, integrin-dependent intravascular neutrophil migration requires CXCR2 ligands on liver sinusoids and formyl peptides in the injury zone[13]. Here we have investigated extravascular neutrophil swarming at very small focal sites of sterile tissue injury in the skin (~1,000× smaller than those examined in the liver model)[4] and in infected lymph nodes[3]. We found that local cell death initiates dramatic swarm-like interstitial neutrophil recruitment and clustering, with a key role for LTB4 as a unique intercellular communication signal between neutrophils that allows rapid integrin-independent neutrophil recruitment through the tissue. Such insights should prove useful for studies in which local sterile or pathogen-induced cell death characterizes innate immune cell dynamics[3–11], and the molecules identified may serve as potential targets for therapeutic intervention in destructive neutrophil-dependent inflammatory processes.

METHODS SUMMARY

Mice. Supplementary Information lists all mouse strains used in this study. All mice were maintained in specific-pathogen-free conditions at an Association for Assessment and Accreditation of Laboratory Animal Care-accredited animal facility at the NIAID and were used under a study protocol approved by NIAID Animal Care and Use Committee (National Institutes of Health).

2P-IVM and focal tissue damage. 2P-IVM of ear pinnae of anaesthetized mice was performed as previously described[5]. For the study of endogenous innate immune cells in the extravascular space, anaesthetized transgenic reporter mice underwent a 15-s skin trauma to recruit neutrophils from the circulation to the dermal interstitium, 3–4 h before focal tissue damage was induced. For gene-function experiments, neutrophils were isolated from bone marrow of control and gene-deficient mice, differentially dye-labelled, and injected intradermally at a 1:1 ratio into *Tyr*$^{c-2J/c-2J}$ mice, 2–3 h before focal tissue damage was induced. Focal tissue damage by a brief two-photon laser pulse (80 mW) has been described previously[4]. All imaged mice were on the *Tyr*$^{c-2J/c-2J}$ (B6.Albino) background to avoid laser-induced cell death of light-sensitive skin melanophages.

Image analysis. Three-dimensional object tracking using Imaris (Bitplane) retrieved cell spatial coordinates over time. These data were further processed using routines developed in the open source programming language R to retrieve dynamic parameters over time for individual cells (distance–time plot) and cell populations (radial velocity–time plot). For details on the source code and definition of chemotactic index, radial velocity and accumulation index, see Supplementary Information. Student's *t*-tests were performed after data were confirmed to fulfil the criteria of normal distribution and equal variance, otherwise Mann–Whitney *U*-tests were applied. Analyses were performed with GraphPad Prism 5 software.

Full Methods and any associated references are available in the online version of the paper.

Received 7 November 2012; accepted 11 April 2013.

Published online 26 May 2013.

1. Nathan, C. Points of control in inflammation. *Nature* **420,** 846–852 (2002).
2. Ley, K., Laudanna, C., Cybulsky, M. I. & Nourshargh, S. Getting to the site of inflammation: the leukocyte adhesion cascade updated. *Nature Rev. Immunol.* **7,** 678–689 (2007).
3. Chtanova, T. *et al.* Dynamics of neutrophil migration in lymph nodes during infection. *Immunity* **29,** 487–496 (2008).
4. Ng, L. G. *et al.* Visualizing the neutrophil response to sterile tissue injury in mouse dermis reveals a three-phase cascade of events. *J. Invest. Dermatol.* **131,** 2058–2068 (2011).
5. Peters, N. C. *et al. In vivo* imaging reveals an essential role for neutrophils in leishmaniasis transmitted by sand flies. *Science* **321,** 970–974 (2008).

6. Bruns, S. *et al.* Production of extracellular traps against *Aspergillus fumigatus in vitro* and in infected lung tissue is dependent on invading neutrophils and influenced by hydrophobin RodA. *PLoS Pathog.* **6,** e1000873 (2010).
7. Yipp, B. G. *et al.* Infection-induced NETosis is a dynamic process involving neutrophil multitasking *in vivo*. *Nature Med.* **81,** 1386–1393 (2012).
8. Liese, J., Rooijakkers, S. H., van Strijp, J. A., Novick, R. P. & Dustin, M. L. Intravital two-photon microscopy of host–pathogen interactions in a mouse model of *Staphylococcus aureus* skin abscess formation. *Cell Microbiol.* http://dx.doi.org/10.1111/cmi.12085 (2012).
9. Harvie, E. A., Green, J. M., Neely, M. N. & Huttenlocher, A. Innate immune response to *Streptococcus iniae* infection in zebrafish larvae. *Infect. Immun.* **81,** 110–121 (2013).
10. Kreisel, D. *et al. In vivo* two-photon imaging reveals monocyte-dependent neutrophil extravasation during pulmonary inflammation. *Proc. Natl Acad. Sci. USA* **107,** 18073–18078 (2010).
11. Nakasone, E. S. *et al.* Imaging tumor-stroma interactions during chemotherapy reveals contributions of the microenvironment to resistance. *Cancer Cell* **21,** 488–503 (2012).
12. Lämmermann, T. *et al.* Rapid leukocyte migration by integrin-independent flowing and squeezing. *Nature* **453,** 51–55 (2008).
13. McDonald, B. *et al.* Intravascular danger signals guide neutrophils to sites of sterile inflammation. *Science* **330,** 362–366 (2010).
14. Guggenberger, C., Wolz, C., Morrissey, J. A. & Heesemann, J. Two distinct coagulase-dependent barriers protect *Staphylococcus aureus* from neutrophils in a three dimensional *in vitro* infection model. *PLoS Pathog.* **8,** e1002434 (2012).
15. McDonald, B. & Kubes, P. Cellular and molecular choreography of neutrophil recruitment to sites of sterile inflammation. *J. Mol. Med.* **89,** 1079–1088 (2011).
16. Sánchez-Madrid, F. & del Pozo, M. A. Leukocyte polarization in cell migration and immune interactions. *EMBO J.* **18,** 501–511 (1999).
17. Cho, H. *et al.* The loss of RGS protein-Gα_{i2} interactions results in markedly impaired mouse neutrophil trafficking to inflammatory sites. *Mol. Cell. Biol.* **32,** 4561–4571 (2012).
18. Kim, N. D., Chou, R. C., Seung, E., Tager, A. M. & Luster, A. D. A unique requirement for the leukotriene B4 receptor BLT1 for neutrophil recruitment in inflammatory arthritis. *J. Exp. Med.* **203,** 829–835 (2006).
19. Chen, M. *et al.* Neutrophil-derived leukotriene B4 is required for inflammatory arthritis. *J. Exp. Med.* **203,** 837–842 (2006).
20. Oyoshi, M. K. *et al.* Leukotriene B4-driven neutrophil recruitment to the skin is essential for allergic skin inflammation. *Immunity* **37,** 747–758 (2012).
21. Chou, R. C. *et al.* Lipid-cytokine-chemokine cascade drives neutrophil recruitment in a murine model of inflammatory arthritis. *Immunity* **33,** 266–278 (2010).
22. Sadik, C. D., Kim, N. D., Iwakura, Y. & Luster, A. D. Neutrophils orchestrate their own recruitment in murine arthritis through C5aR and FcγR signaling. *Proc. Natl Acad. Sci. USA* **109,** E3177–E3185 (2012).
23. Afonso, P. V. *et al.* LTB4 is a signal-relay molecule during neutrophil chemotaxis. *Dev. Cell* **22,** 1079–1091 (2012).
24. DiPersio, J. F., Billing, P., Williams, R. & Gasson, J. C. Human granulocyte-macrophage colony-stimulating factor and other cytokines prime human neutrophils for enhanced arachidonic acid release and leukotriene B4 synthesis. *J. Immunol.* **140,** 4315–4322 (1988).
25. Malawista, S. E., de Boisfleury Chevance, A., van Damme, J. & Serhan, C. N. Tonic inhibition of chemotaxis in human plasma. *Proc. Natl Acad. Sci. USA* **105,** 17949–17954 (2008).
26. Palmblad, J. *et al.* Leukotriene B4 is a potent and stereospecific stimulator of neutrophil chemotaxis and adherence. *Blood* **58,** 658–661 (1981).
27. Ford-Hutchinson, A. W., Bray, M. A., Doig, M. V., Shipley, M. E. & Smith, M. J. Leukotriene B, a potent chemokinetic and aggregating substance released from polymorphonuclear leukocytes. *Nature* **286,** 264–265 (1980).
28. Calderwood, D. A. & Ginsberg, M. H. Talin forges the links between integrins and actin. *Nature Cell Biol.* **5,** 694–697 (2003).
29. Kastenmüller, W., Torabi-Parizi, P., Subramanian, N., Lämmermann, T. & Germain, R. N. A spatially-organized multicellular innate immune response in lymph nodes limits systemic pathogen spread. *Cell* **150,** 1235–1248 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Birnbaumer, R. Fässler, E. Tuomanen, P. M. Murphy, S. Akira, S. Monkley, D. Critchley, R. Wedlich-Söldner and M. Sixt for providing mice for this study, J.G. Egen and J. Tang for assistance with imaging, M. Parsek for providing *P. aeruginosa*–GFP, J.H. Kehrl, P.M. Murphy, R. Varma and members of the Germain laboratory for discussions. This work was supported by the Intramural Research Program of National Institute of Allergy and Infectious Diseases and National Cancer Institute, National Institutes of Health. T.L. was supported by a Human Frontiers Science Program Long-Term Fellowship. W.K. is presently a member of the Deutsche Forschungsgemeinschaft (DFG)-funded excellence cluster ImmunoSensation, Bonn, Germany.

Author Contributions T.L. and R.N.G. designed the experiments, interpreted the data and wrote the paper. P.V.A., J.M.W. and C.A.P. contributed to data interpretation and experimental design, as well as the development of the final version of the paper. T.L. performed all skin-imaging experiments. W.K. performed intravital imaging of infected lymph nodes. B.R.A. developed the software for graphical display of imaging data. T.L., B.R.A. and P.V.A. conducted quantitative analysis of the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.L. (laemmermannt@niaid.nih.gov) or R.N.G. (rgermain@niaid.nih.gov).

METHODS

Mice. Supplementary Table 1 lists all mouse strains and crosses used in this study. $Gnai2^{-/-}$[30], $Gnai3^{-/-}$[31], $Itgb1^{fl/fl}$[32], $Tln1^{fl/fl}$[33], $Ccr1^{-/-}$[34], $Fpr1^{-/-}$[35], $Fpr2^{-/-}$[36], $Ptafr^{-/-}$[37], $Myd88^{-/-}$[38] and Lifeact–$GFP^{+/-}$[39] mice have been described elsewhere. $Lyz2^{gfp/gfp}$[40] and $Il1r1^{-/-}$[41] mice were obtained from Taconic Laboratories through a special contract with the NIAID. All other mouse strains were purchased from Jackson Laboratories. All mice were maintained in specific-pathogen-free conditions at an Association for Assessment and Accreditation of Laboratory Animal Care-accredited animal facility at the NIAID and were used under a study protocol approved by NIAID Animal Care and Use Committee (National Institutes of Health).

2P-IVM of ear skin and infected lymph nodes. Two-photon intravital imaging of ear pinnae of anaesthetized mice was performed as previously described[5,42]. Mice were anaesthetized using isoflurane (Baxter; 2% for induction, 1–1.5% for maintenance, vaporized in an 80:20 mixture of oxygen and air) and placed in a lateral recumbent position on a custom imaging platform such that the ventral side of the ear pinna rested on a coverslip. A strip of Durapore tape was placed lightly over the ear pinna and affixed to the imaging platform to immobilize the tissue. Care was taken to minimize pressure on the ear. Images were captured towards the anterior half of the ear pinna where hair follicles are sparse. Images were acquired using an inverted LSM 510 NLO multiphoton microscope (Carl Zeiss Microimaging) enclosed in a custom-built environmental chamber that was maintained at 32 °C using heated air. This system had been custom fitted with three external non-descanned photomultiplier tube detectors in the reflected light path. Images were acquired using a 25×/0.8 numerical aperture (NA) Plan-Apochromat objective (Carl Zeiss Imaging) with glycerol as immersion medium. Fluorescence excitation was provided by a Chameleon XR Ti:Sapphire laser (Coherent) tuned to 850 nm for dye excitation and generation of collagen second harmonic signal, or 920 nm for enhanced GFP (eGFP) excitation and 940 nm for excitation of both DsRed and eGFP. For four-dimensional data sets, three-dimensional stacks were captured every 30 s, unless otherwise specified. All imaged mice were on the $Tyr^{c-2J/c-2J}$ (B6.Albino) background to avoid laser-induced cell death of light-sensitive skin melanophages[43]. 2P-IVM on *P. aeruginosa*-infected lymph nodes was performed as previously described[29]. 10^7 colony-forming units of GFP-expressing *Pseudomonas aeruginosa*[44] (provided by M. Parsek) were diluted in PBS and injected in the footpad (30 µl) of wild-type $Lyz2^{gfp/+}$ and $Ltb4r1^{-/-}$ $Lyz2^{gfp/+}$ mice and draining lymph nodes imaged at indicated times after infection using a Zeiss 710 microscope equipped with a Chameleon laser (Coherent) and a 20× water-dipping lens (NA 1.0, Zeiss). The microscope was enclosed in an environmental chamber in which anaesthetized mice were warmed by heated air, and the surgically exposed lymph node was kept at 36–37 °C with warmed PBS. For 2P-IVM of the subcapsular sinus, we used a *z*-stack of 40–50 µm, 3-µm step size and acquired images every 40 s. Raw imaging data were processed with Imaris (Bitplane) using a Gaussian filter for noise reduction. All images and movies are displayed as two-dimensional maximum-intensity projections of 10–30-µm-thick *z*-stacks.

Endogenous mouse innate immune cells. For the study of endogenous innate immune cells in the extravascular space, transgenic reporter mice were crossed with $Tyr^{c-2J/c-2J}$ (B6.Albino) mice to yield $Lyz2^{gfp/+}$ $Tyr^{c-2J/c-2J}$ (myelomonocytic cells), $DsRed^{+/-}$ $Lyz2^{gfp/+}$ $Tyr^{c-2J/c-2J}$ (myelomonocytic cells, stroma) and $DsRed^{+/+}$ $Cx3cr1^{gfp/gfp}$ $Tyr^{c-2J/c-2J}$ (macrophages/monocytes, stroma and neutrophils) mice. These animals were anaesthetized with isoflurane and underwent a brief skin trauma to recruit neutrophils from the circulation to the dermal interstitium. The anaesthetized mouse was placed on a scale and 30 N per cm^2 pressure was applied for 15–20 s on the mouse ear with the investigator's thumb. As this method induced initial neutrophil extravasation that stopped after 2–3 h, this method was superior to other tested common inflammatory treatments (for example, chemical skin irritation) that all lead to neutrophil recruitment from blood over longer time periods. Three hours after brief skin trauma, mice were prepared for skin imaging as described above and rested in the heated environmental chamber for 30–60 min before the first focal tissue damage was induced.

Neutrophil isolation, labelling and i.d. injection. For i.d. injection experiments, mouse neutrophils were isolated from bone marrow using a three-layer Percoll gradient of 78%, 69% and 52% as previously described[45]. Neutrophils were collected at the 69–78% interface and were highly purified (>98%) as indicated by Lyz2–$eGFP^{hi}$ $Ly6G^{pos}$ $Ly6C^{neg}$ phenotype in flow cytometry[46] (data not shown), with viability >98% by trypan blue staining. Neutrophils were washed three times with washing buffer (1× Hank's balanced salt solution, 1% FBS, 2 mM EDTA). If further labelled with cell dyes, neutrophils were incubated for 15 min with either 0.8 µM cell tracker red (CMTPX) or 1 µM cell tracker green (CMFDA) in 1× HBSS supplemented with 0.0002% (w/v) pluronic F-127 (all Life Technologies). Neutrophils were washed four times with washing buffer, before a 1:1 ratio of differentially labelled control and gene-deficient neutrophils (each $>2 \times 10^6$ cells) was taken up in 1× PBS at a volume of 15–30 µl. A volume of 5 µl neutrophil suspension was injected intradermally with an insulin syringe (31.5 GA needle, BD Biosciences) into the ventral side of the mouse ear pinnae. Recipient mice were always on the $Tyr^{c-2J/c-2J}$ (B6.Albino) background. Two hours after injection, mice were prepared for skin imaging as described above and rested in the heated environmental chamber for 30–60 min before the first focal tissue damage was induced. For isolation of talin-deficient neutrophils, $Tln1^{fl/fl}$ mice were intercrossed with $Mx1\text{-}cre^{+/-}$ mice[47] to yield $Tln1^{fl/fl}$ $Mx1\text{-}cre^{+/-}$. In these mice, Cre expression in the haematopoietic system was induced by intraperitoneal injection of 250 µg Poly(I):Poly(C) (Amersham Biosciences). Five days after knockout induction neutrophils were isolated from bone marrow with high efficiency of talin depletion[12]. For most injection experiments, we interchanged dyes to exclude unspecific effects and never observed differences in migration or accumulation phenotypes. Whereas cell tracker green labels neutrophils homogeneously, cell tracker red gives a polarized neutrophil staining over time. Cell tracker red also leaks over time from cells, resulting in background fluorescence in the tissue and some fluorescent resident, non-motile, elongated macrophages taking up the dye. For experiments with human neutrophils, heparinized whole blood was obtained by venipuncture from healthy donors. Blood samples were obtained from anonymous blood donors enrolled in the NIH Blood Bank research program. Human neutrophils were isolated with Lympholyte-poly Cell Separation Media (Cedarline) as previously described[48]. Residual erythrocytes were removed using ammonium-chloride-potassium lysis buffer (Lonza).

Focal tissue damage. A protocol for focal skin tissue damage by a focused two-photon laser pulse has been described previously[43] and used with slight modifications. The Chameleon XR Ti:sapphire laser (Coherent) was tuned to 850 nm and the laser intensity adjusted to 80 mW. At pixel dimensions of 0.14 × 0.14 µm, a circular region of interest of 25–35 µm in diameter (approximately $1\text{–}2 \times 10^{-6}$ mm^3 in volume) (unless otherwise specified) was defined in one focal plane, followed by laser scanning at a pixel dwell time of 0.8 µs for 35–50 iterations, depending on the tissue depth of the imaging field of view. The damage was restricted to dermal layers only (Supplementary Fig. 2). Immediately after laser-induced tissue damage, imaging of the neutrophil response was started at typical voxel dimensions of 0.72 × 0.72 × 2 µm. We performed a maximum of three consecutive experiments per mouse ear.

Data analysis. Three-dimensional object tracking using Imaris (Bitplane) retrieved cell spatial coordinates (x, y, z) over time. These data were further processed using routines composed in the open source programming language R (source code in Supplementary Material) to retrieve dynamic parameters for individual cells (distance-time plot, DTP) and cell populations (radial velocity–time plot). The chemotactic index was defined as cos(α), with α as the angle between the distance vector to the damage site and the actual movement vector (Supplementary Fig. 4). The instantaneous radial velocity is the product of the chemotactic index and the cell's instantaneous velocity (migrated distance over 30 s). Reconstructed cell tracks were filtered using a moving average with a width of 2.5 min to suppress spurious direction changes due to limited resolution of the tracking algorithm and small scale features such as other cells in the path of the migrating cell. Paths that had durations of less than 3 min were dropped. Velocities and the chemotactic index were calculated using central differences from the filtered paths. Plots including regression lines were generated using the ggplot2 package (version 0.8.9) for R (version 2.13.1).

The accumulation index as measure of cell entry into the collagen-free zone was defined as the ratio of fluorescent signal from gene-deficient cells in the collagen-free zone versus total signal at the wound site divided by the ratio of fluorescent signal from control cells in the collagen-free zone versus total signal at the wound site. Fluorescent signals from static 2P-IVM images were quantified in ImageJ software (National Institutes of Health).

Immunofluorescence in ear skin whole mounts. One to two hours after induction of several focal tissue damage sites, mice were euthanized, ears excised, divided into dorsal and ventral halves, ventral halves fixed in 1% paraformaldehyde (Electron Microscopy Sciences) and stained with antibodies diluted in washing buffer consisting of 1× PBS, 1% BSA and 0.025% (v/v) Triton X-100 (Sigma-Aldrich). For identification of neutrophils in $DsRed^{+/+}$ $Cx3cr1^{gfp/gfp}$ $Tyr^{c-2J/c-2J}$ mice, ventral ear whole mounts were stained with Alexa Fluor 647 anti-mouse Ly-6G Antibody (clone 1A8, BioLegend)[49]. For detection of CXCL2 (MIP-2) and CRAMP in neutrophil clusters, anti-MIP2 antibody (R&D Systems), anti-CRAMP antibody (Phoenix Pharmaceuticals), rabbit IgG isotype control and goat IgG isotype control (both Southern Biotech) were all conjugated to Alexa Fluor 555 according to the manufacturer's instructions (Life Technologies). Fixed ventral ear halves of $Lyz2^{gfp/gfp}$ mice were blocked with 10% normal goat serum or normal rabbit serum (both Southern Biotech) in washing buffer before staining with the Alexa Fluor 555-conjugated antibodies. All Alexa Fluor 555-conjugated antibodies gave strong punctate staining in the dermal tissue, but isotype controls never gave staining within neutrophils. Confocal microscopy on skin whole mounts was performed with a LSM 710 confocal microscope equipped with a 20×/0.8 NA Plan-Apochromat objective (Carl Zeiss Microimaging) using 1-μm optical slices.

Statistical analysis. Student's *t*-tests were performed after data were confirmed to fulfil the criteria of normal distribution and equal variance; otherwise Mann–Whitney *U*-tests were applied. Analyses were performed with GraphPad Prism 5 software.

30. Rudolph, U. *et al.* Ulcerative colitis and adenocarcinoma of the colon in Gα_{i2}-deficient mice. *Nature Genet.* **10,** 143–150 (1995).
31. Pero, R. S. *et al.* Gα_{i2}-mediated signaling events in the endothelium are involved in controlling leukocyte extravasation. *Proc. Natl Acad. Sci. USA* **104,** 4371–4376 (2007).
32. Potocnik, A. J., Brakebusch, C. & Fassler, R. Fetal and adult hematopoietic stem cells require beta1 integrin function for colonizing fetal liver, spleen, and bone marrow. *Immunity* **12,** 653–663 (2000).
33. Petrich, B. G. *et al.* Talin is required for integrin-mediated platelet function in hemostasis and thrombosis. *J. Exp. Med.* **204,** 3103–3111 (2007).
34. Gao, J. L. *et al.* Impaired host defense, hematopoiesis, granulomatous inflammation and type 1-type 2 cytokine balance in mice lacking CC chemokine receptor 1. *J. Exp. Med.* **185,** 1959–1968 (1997).
35. Gao, J. L., Lee, E. J. & Murphy, P. M. Impaired antibacterial host defense in mice lacking the *N*-formylpeptide receptor. *J. Exp. Med.* **189,** 657–662 (1999).
36. Chen, K. *et al.* A critical role for the G protein-coupled receptor mFPR2 in airway inflammation and immune responses. *J. Immunol.* **184,** 3331–3335 (2010).
37. Radin, J. N. *et al.* β-Arrestin 1 participates in platelet-activating factor receptor-mediated endocytosis of *Streptococcus pneumoniae*. *Infect. Immun.* **73,** 7827–7835 (2005).
38. Adachi, O. *et al.* Targeted disruption of the MyD88 gene results in loss of IL-1- and IL-18-mediated function. *Immunity* **9,** 143–150 (1998).
39. Riedl, J. *et al.* Lifeact mice for studying F-actin dynamics. *Nature Methods* **7,** 168–169 (2010).
40. Faust, N., Varas, F., Kelly, L. M., Heck, S. & Graf, T. Insertion of enhanced green fluorescent protein into the lysozyme gene creates mice with green fluorescent granulocytes and macrophages. *Blood* **96,** 719–726 (2000).
41. Glaccum, M. B. *et al.* Phenotypic and functional characterization of mice that lack the type I receptor for IL-1. *J. Immunol.* **159,** 3364–3371 (1997).
42. Gaiser, M. R. *et al.* Cancer-associated epithelial cell adhesion molecule (EpCAM; CD326) enables epidermal Langerhans cell motility and migration *in vivo*. *Proc. Natl Acad. Sci. USA* **109,** E889–E897 (2012).
43. Li, J. L. *et al.* Intravital multiphoton imaging of immune responses in the mouse ear skin. *Nature Protocols* **7,** 221–234 (2012).
44. Davies, D. G. *et al.* The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science* **280,** 295–298 (1998).
45. Boxio, R., Bossenmeyer-Pourie, C., Steinckwich, N., Dournon, C. & Nusse, O. Mouse bone marrow contains large numbers of functionally competent neutrophils. *J. Leukoc. Biol.* **75,** 604–611 (2004).
46. Woodfin, A. *et al.* The junctional adhesion molecule JAM-C regulates polarized transendothelial migration of neutrophils *in vivo*. *Nature Immunol.* **12,** 761–769 (2011).
47. Kühn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science* **269,** 1427–1429 (1995).
48. Oh, H., Siano, B. & Diamond, S. Neutrophil isolation protocol. *J. Vis. Exp.* **17,** 745 (2008).
49. Pflicke, H. & Sixt, M. Preformed portals facilitate dendritic cell entry into afferent lymphatic vessels. *J. Exp. Med.* **206,** 2925–2935 (2009).

doi:10.1038/nature12274

HIV-1 causes CD4 cell death through DNA-dependent protein kinase during viral integration

Arik Cooper[1], Mayra García[1], Constantinos Petrovas[2], Takuya Yamamoto[2], Richard A. Koup[2] & Gary J. Nabel[1]†

Human immunodeficiency virus-1 (HIV-1) has infected more than 60 million people and caused nearly 30 million deaths worldwide[1], ultimately the consequence of cytolytic infection of CD4$^+$ T cells. In humans and in macaque models, most of these cells contain viral DNA and are rapidly eliminated at the peak of viraemia[2–4], yet the mechanism by which HIV-1 induces helper T-cell death has not been defined. Here we show that virus-induced cell killing is triggered by viral integration. Infection by wild-type HIV-1, but not an integrase-deficient mutant, induced the death of activated primary CD4 lymphocytes. Similarly, raltegravir, a pharmacologic integrase inhibitor, abolished HIV-1-induced cell killing both in cell culture and in CD4$^+$ T cells from acutely infected subjects. The mechanism of killing during viral integration involved the activation of DNA-dependent protein kinase (DNA-PK), a central integrator of the DNA damage response, which caused phosphorylation of p53 and histone H2AX. Pharmacological inhibition of DNA-PK abolished cell death during HIV-1 infection *in vitro*, suggesting that processes which reduce DNA-PK activation in CD4 cells could facilitate the formation of latently infected cells that give rise to reservoirs *in vivo*. We propose that activation of DNA-PK during viral integration has a central role in CD4$^+$ T-cell depletion, raising the possibility that integrase inhibitors and interventions directed towards DNA-PK may improve T-cell survival and immune function in infected individuals.

To investigate the molecular mechanism underlying HIV-1-induced cell death during acute infection, we first infected CEMX174 T leukaemia cells with HIV-1$_{\mathrm{NL4\text{-}3}}$ and monitored viral replication and cell death (Fig. 1a). Cells expressing p24 (p24$^+$) remained viable throughout the experiment. By contrast, the majority of cells lacking p24 (p24$^-$) were nonviable by day 6. Similar results were obtained using a single cycle, green fluorescent protein (GFP) encoding vesicular stomatitis virus-G (VSVG)-pseudotyped lentiviral vector (Fig. 1b), indicating that death was independent of the viral envelope and syncytia formation and occurred during a single cycle of viral replication. Cell death induced by this vector was more rapid and correlated with the faster kinetics of infection conferred by VSVG (ref. 5). Primary human CD4 cells showed a similar effect, whether infected by laboratory or primary virus isolates (Fig. 1c and Supplementary Fig. 1).

To determine whether cells lacking viral gene expression had been productively infected before cell death, primary lymphocytes were infected with a replication-competent HIV-1 encoding GFP and sorted for expression and viability (Fig. 2a). Viral cDNA was detected by quantitative real-time PCR (qPCR) in nonviable GFP-negative cells, indicating that these cells had been infected (Fig. 2b). Time-course analysis of viable GFP-positive cells sorted from a culture infected with a GFP-encoding HIV-1 revealed that a very high proportion of these cells died within 2 days and lost GFP expression, yet they retained viral DNA (Fig. 2c and Supplementary Fig. 2a), indicating the dying cells had undergone productive infection that enabled transient reporter gene expression before causing cell death.

To determine which step in viral replication was responsible for cell death, primary lymphocytes were infected in the absence or presence of raltegravir, an integrase inhibitor. Indinavir, a viral protease inhibitor that blocks HIV-1 egress by preventing capsid maturation, was included to restrict viral infection to a single round. In the absence of raltegravir, p24$^-$ cells from the infected culture were largely nonviable (Fig. 3a), as shown above. Raltegravir diminished both cell death and proviral DNA integration. The level of two long terminal repeat (2 LTR) circles increased in the presence of raltegravir, probably because unintegrated reverse transcripts circularize in cells treated with this drug[6]. Efavirenz, a non-nucleoside reverse transcriptase inhibitor that acts earlier in the virus life cycle, also inhibited cell death (Fig. 3a), as expected. Similarly, both raltegravir and efavirenz diminished death of CEMX174 cells infected with a VSVG lentiviral vector (Supplementary Fig. 2b). AZT (3′-azido-3′-deoxythymidine), a late-stage reverse transcriptase inhibitor, and D-118-24, an integrase inhibitor with similar activity to that of raltegravir, also each prevented cell death (Supplementary Fig. 2b).

To document this point, we infected primary CD4 lymphocytes or CEMX174 cells with VSVG-pseudotyped HIV-1 containing a D64V point mutation that abolishes viral DNA integration[7]. Infection with this integrase mutant virus failed to trigger death in either cell type (Fig. 3b and Supplementary Fig. 2c). qPCR confirmed the near complete absence of provirus integration and the accumulation of 2 LTR circles in cells infected with this mutant (Fig. 3b). Together, these results demonstrated that integration was required for death during afferent HIV-1 infection.

Provirus integration requires nuclear import of fully reverse-transcribed viral genomes followed by strand transfer into the host chromosome. Notably, the DNA ends of reverse transcripts may promote apoptosis in cells with defective non-homologous DNA end joining[8]. Because nuclear reverse transcripts form 2 LTR circles in the absence of chromosomal integration[6], inhibition of cell death by a mutant integrase or by raltegravir could result either from cDNA circularization or from a block to strand transfer. To distinguish between these alternatives, we performed knockdown of DNA ligase 4, an enzyme required for 2 LTR circle formation[8]. This treatment substantially reduced 2 LTR circles and concomitantly increased unintegrated reverse transcripts in the nuclei of cells infected with an integrase mutant virus (Fig. 3c and Supplementary Fig. 2d). HIV-1-induced cell death was nonetheless blocked under these conditions (Fig. 3c), suggesting that DNA circularization is not involved in this pathway, and that death signalling requires strand transfer into the host chromosome.

To test whether proviral integration was sufficient to cause cell killing, we used a *tat*- and *rev*-deficient HIV-1 that eliminates viral gene expression without affecting the early steps including integration[9]. This mutant integrated similarly to a matching wild-type virus, and although no viral gene expression was detected, it caused substantial cell death (Fig. 3d). Integration was therefore both necessary and sufficient for triggering cell death by HIV-1.

[1]Virology Laboratory, Vaccine Research Center, National Institute for Allergy and Infectious Diseases, National Institutes of Health, Building 40, Room 4502, MSC-3005, 40 Convent Drive, Bethesda, Maryland 20892-3005, USA. [2]Immunology Laboratory, Vaccine Research Center, National Institute for Allergy and Infectious Diseases, National Institutes of Health, Building 40, Room 3502, MSC-3022, 40 Convent Drive, Bethesda, Maryland 20892-3022, USA. †Present address: Sanofi, 640 Memorial Drive, Cambridge, Massachusetts 02139, USA.

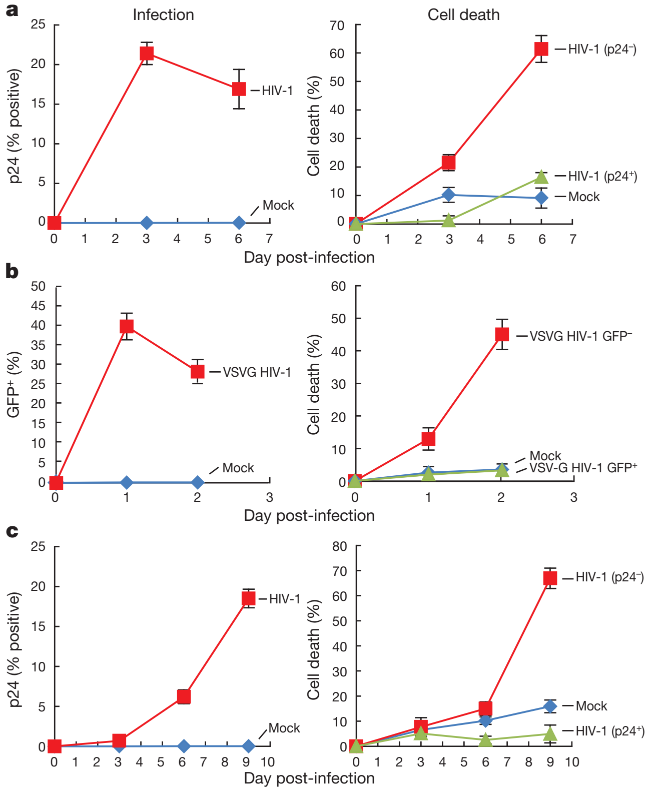


Figure 1 | CD4 lymphocytes killed during HIV-1 infection do not express viral gene products. Time-course analysis of viral replication and cell viability in CEMX174 cells infected with replication-competent HIV-1_{NL4-3} (**a**) or a single-cycle HIV-1 vector encoding GFP (VSVG HIV-1) (**b**) and in primary CD4^{+} T cells infected with HIV-1_{NL4-3} (**c**). Viral replication was monitored using intracellular p24 Gag staining and cell viability analysis was simultaneously performed on gated p24^{-} and p24^{+} populations using Annexin V and the permeability dye Vivid. Time points are indicated at the bottom of each panel. Data are representative of four independent experiments and conducted in triplicate. Error bars represent the standard deviation of the mean for each time point.

We examined this mechanism in natural infection using peripheral blood mononuclear cells from untreated infected subjects. These cells were activated *ex vivo* to induce viral replication, yielding CD3^{hi} CD8^{-} p24^{+} cells (Fig. 3e, left and Supplementary Fig. 3a) that showed a statistically significantly reduction in death compared with CD4^{hi} p24^{-} cells (Fig. 3e, middle and Supplementary Fig. 3b), similar to the *in vitro* experiments. Furthermore, dying p24^{-} CD4 lymphocytes contained viral DNA (Fig. 3e, right). Raltegravir substantially restored viability under these conditions (Supplementary Fig. 3c, d), indicating that viral integration significantly contributed to cell death. Together these findings suggest that cell death arises from a similar mechanism of signaling during natural infection.

We further explored the cell death mechanism by analysing markers of the double-stranded DNA damage response as they relate to the final steps of viral integration. Infection by both replication-competent HIV-1 and a single-cycle lentiviral vector stimulated DNA-PK activity and phosphorylation of p53 in primary CD4 lymphocytes and CEMX174 cells, as well as triggering H2AX phosphorylation (Fig. 4a). This response was observed with an integrase-competent VSVG HIV-1 but not with a matching integrase-mutant virus (Fig. 4a, right panel). DNA-PK localized to the nuclei of infected cells, confirming the presence of both the kinase and the viral genome in the relevant cellular compartment (Supplementary Fig. 4a). Together these results reveal that integration elicits a cellular double-stranded DNA damage response, suggesting a causal link between DNA-PK activation, p53 phosphorylation and virus-induced cell killing.

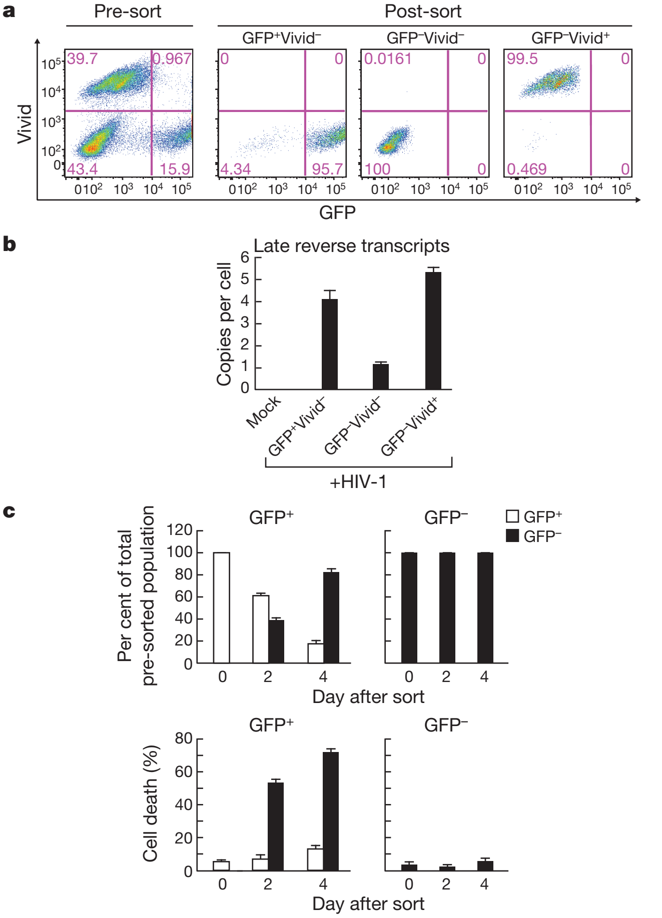


Figure 2 | Dying CD4 lymphocytes lacking viral gene expression had been productively infected prior to cell death. **a**, Flow cytometry analyses of Vivid-stained, primary CD4 lymphocytes infected for 8 days with a GFP-encoding, replication-competent HIV-1 before and after sorting of the indicated populations. **b**, qPCR analysis for late reverse transcripts in the indicated sorted populations shown in **a**. **c**, Time-course analysis of GFP fluorescence and viability in the indicated primary CD4 lymphocyte populations sorted 36 h after infection with VSVG-pseudotyped HIV-1 encoding GFP. Data are representative of two independent experiments from two different donors and conducted in triplicate. Error bars represent the standard deviation of the mean.

We examined the role of DNA-PK in cell death by using NU7026, a pharmacological inhibitor of this kinase. NU7026 abolished p53 and H2AX phosphorylation and blocked death of HIV-1 infected cells without substantially affecting viral replication (Fig. 4b–d and Supplementary Fig. 4b–d). A similar effect was observed with an independent DNA-PK inhibitor, NU7441 (Fig. 4c, d and Supplementary Fig. 4b–d). By contrast, both DNA-PK inhibitors enhanced etoposide-induced killing in primary CD4 lymphocytes (Supplementary Fig. 4e), ruling out non-specific protective effects of these drugs. We further examined the causal link between p53 phosphorylation and cell death using pifithrin, a specific pharmacologic p53 antagonist. This inhibitor substantially decreased cell death following infection of CEMX174 cells (Supplementary Fig. 4f). Finally, ATM, a kinase involved in repair and apoptosis following DNA damage, is also stimulated following infection, but its activation was independent of DNA-PK (Supplementary Fig. 4g). A specific inhibitor of this kinase, KU55933, did not substantially affect the HIV-1-induced damage response and cell death (Fig. 4b and Supplementary Fig. 4h), demonstrating that DNA-PK plays a non-redundant role in inducing cell death after infection.

We demonstrate here that integration triggers signal transduction that causes HIV-1-induced killing in activated CD4 lymphocytes. This cell death occurs in nearly all infected cells and is associated with

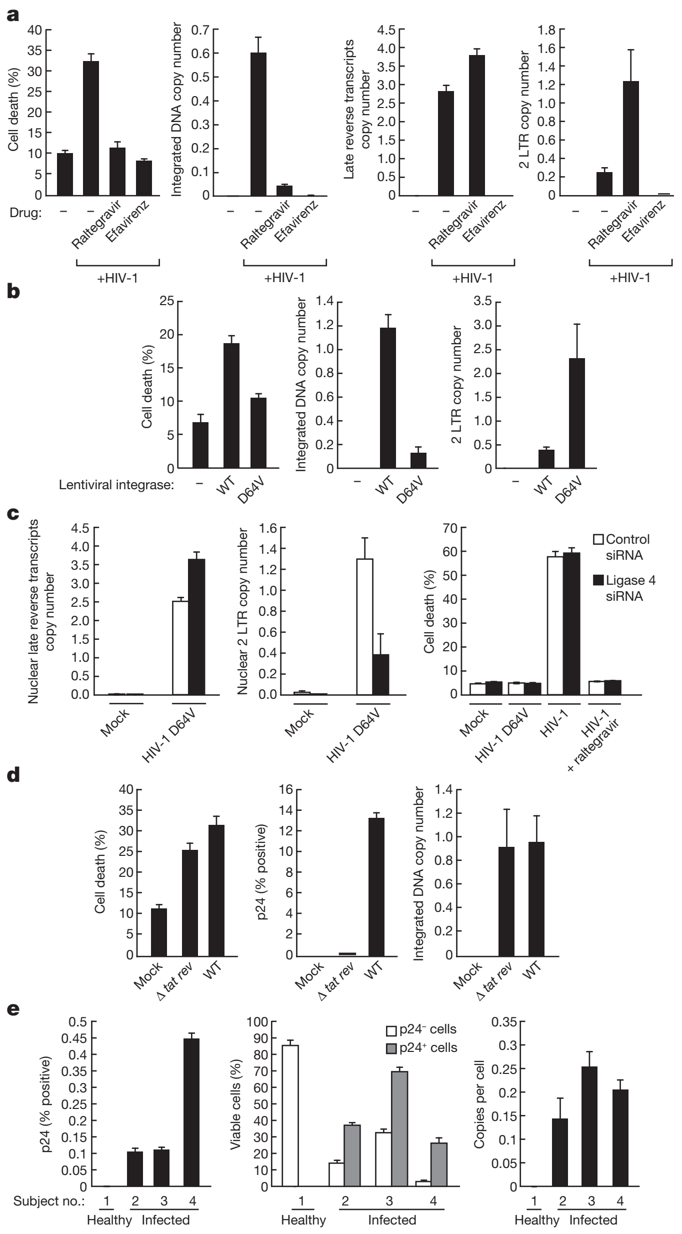


Figure 3 | Proviral DNA integration triggers cell death during HIV-1 infection. **a**, Cell death and the indicated viral DNA analyses in primary CD4 lymphocytes infected with HIV-1_{NL4-3} at a multiplicity of infection (m.o.i) = 1 for 3 days in the presence of indinavir, and in the absence or presence of the indicated inhibitors. **b**, Cell death and viral DNA analyses of primary lymphocytes infected with either a wild-type or a D64V integrase mutant HIV-1 GFP reporter virus. **c**, Cell death and the indicated nuclear viral DNA analyses of cells nucleofected with control or DNA ligase 4 siRNA and infected with the indicated viruses for 3 days. **d**, Cell death, intracellular p24 Gag staining and integrated proviral DNA analyses of primary CD4 lymphocytes either uninfected or infected with the indicated viruses for 3 days in the presence of indinavir. All data are representative of three different experiments done in triplicate. **e**, Death of CD4 lymphocytes lacking p24 expression freshly isolated from HAART-naive HIV-1-infected subjects, and presence of viral DNA in $p24^{-}$ cells. Viral replication was evaluated by intracellular p24 Gag staining in peripheral blood mononuclear cells from a healthy donor (subject 1) and three different viraemic, untreated patients (subjects 2, 3 and 4) that were activated for 3 days (left). Percentages represent the fraction of $CD3^{hi}$ $CD8^{-}$ cells. Cell death was determined using Vivid and Annexin V staining followed by flow cytometry analysis of $CD3^{hi}$ $CD8^{-}$ cells either expressing or lacking intracellular p24 staining in the same samples (middle), and qPCR analysis for HIV-1 late reverse transcripts was performed in dying, $p24^{-}$ $CD4^{+}$ T cells sorted from the same peripheral blood mononuclear cell samples (right). Data are representative of two experiments done in triplicate. Error bars represent the standard deviation of the mean.

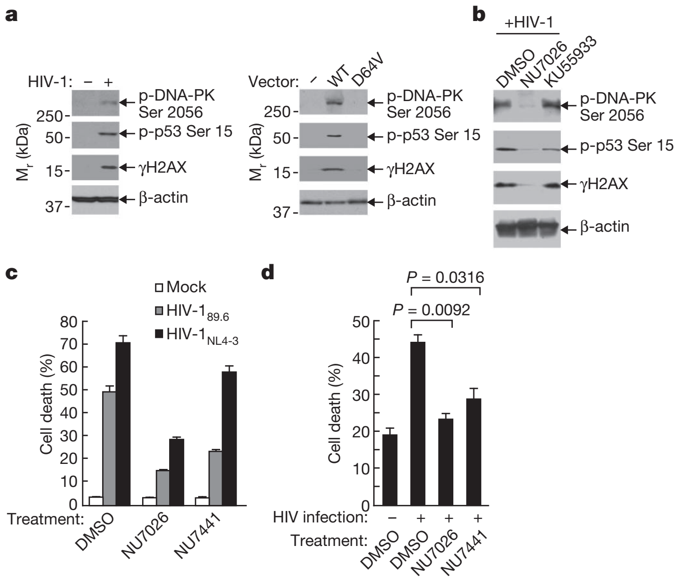


Figure 4 | DNA-PK orchestrates a DNA damage response and cell death following proviral DNA integration. **a**, Western blot analysis showing phosphorylation of DNA-PK, p53 and H2AX following high m.o.i. infections with HIV-1_{NL4-3} in primary CD4 lymphocytes (left panel) or following infection with either VSVG HIV-1 harbouring a wild type (WT) or a D64V integrase mutant (D64V). Extracts from the indicated cultures were analysed 48 h after infection, and cell death peaked after an additional day in the same cultures (data not shown). **b**, Western blot analysis showing phosphorylation of DNA-PK, p53 and H2AX following infection with HIV-1_{NL4-3} in the absence or presence of the indicated DNA-PK and ATM inhibitors. **c**, Cell death analysis of CEMX174 cultures, gated on $p24^{-}$ cells, that were either uninfected or infected with HIV-$1_{89.6}$ or HIV-1_{NL4-3} in the absence or presence of NU7026 or NU7441 for 4 days. **d**, Cell death analysis gated on $p24^{-}$ cells of primary CD4 lymphocyte cultures either uninfected or infected with HIV-$1_{89.6}$ in the absence or presence of NU7026 or NU7441 for 7 days. Data in panels **c** and **d** are representative of at least two independent experiments done in triplicate. Error bars represent the standard deviation of the mean of cell death.

productive, not abortive, infection as previously suggested[10], the differences possibly explained by alternative activation states of the cells and by our finding that HIV-1 gene expression is quickly diminished concomitant with death of productively infected cells. We further demonstrate that DNA-PK is activated by the integrating provirus and conveys the death signal. DNA-PK has been the subject of previous studies of HIV-1 infection, in which its role in the virus life cycle has generated controversy[11–14]. However, none of these studies analysed its role in primary human CD4 lymphocytes, the physiological target of HIV-1 infection. Although we observed no substantial change in viral replication when DNA-PK activity was inhibited in activated T cells, it nonetheless remains possible that DNA-PK plays a non-catalytic role in post-integration repair in some circumstances[11,15].

Lentiviral integration proceeds through a combination of cleavage and ligation such that double-stranded breaks in the host genome are avoided[16]. However, the integrase-catalysed reaction generates short gaps in the host genome at each newly formed junction that can potentially lead to DNA-PK activation directly[17], though cell death was not examined previously. Also, because these gaps are susceptible to breakage, a likely stimulus for DNA-PK activation is the presence of double-stranded DNA breaks which can arise secondarily at these junctions. Binding of DNA-PK subunits to the incoming viral pre-integration complex has been previously reported[8,18] and may prime DNA-PK as the sensor for integrating proviral DNA. There is increasing evidence that this kinase is involved in apoptotic signalling[15,19], and the mechanism controlling the relative DNA-PK activities in damage repair versus cell death involves differential autophosphorylation of this kinase and its subcellular localization during the cell cycle[20,21].

Notably, CD4 T-cell activation, which is strongly associated with HIV-induced depletion *in vivo*[2], promotes nuclear translocation and activation of DNA-PK[22], thus providing further support for its role in activated T cells. This observation further indicates that DNA-PK may respond differently during infection in other cell types, such as macrophages or even resting CD4 lymphocytes, a context in which HIV-1 replication may be less cytotoxic[23–25], potentially facilitating proviral integration and establishment of the latent reservoir.

Neither the role of viral integrase nor its effect on DNA-PK has been implicated previously in HIV-1-induced primary CD4 lymphocyte death, and thus this study defines a specific mechanism that significantly contributes to immunodeficiency caused by HIV-1. It may also provide an explanation for the short half-life and high turnover of productively infected CD4 lymphocytes documented in infected individuals[26,27]. Finally, it remains possible that antiviral therapy with integrase and/or DNA-PK inhibitors can help preserve T-cell function *in vivo*. This possibility has been raised recently by a study in patients switching to a raltegravir-containing regimen[28], and it therefore deserves further scrutiny.

METHODS SUMMARY

CD4 T cells were isolated from elutriated lymphocytes of healthy donors using magnetic bead purification, activated and infected as described in the Methods section. Samples from HIV-1-infected subjects were thawed and passed through a dead cell capture column (Miltenyi Biotech) before cell activation. Where indicated, antiretroviral drugs or DNA-PK inhibitors were used at the concentrations given in the Methods. Viability of infected cultures was determined using combined staining with Annexin V-APC (BD Pharmingen) and the amine reactive viability dye Vivid (Invitrogen). In some experiments, the cells were fixed, permeabilized and stained for intracellular p24 Gag (FITC (fluorescein isothiocyanate) or KC-57 PE; Coulter) before flow cytometry analysis with a LSR II cell analyser. Sorting of live, infected cells was done in a BSL-3 facility using a FACSAria sorter. For viral DNA analysis, the DNeasy Blood and Tissue Kit (Qiagen) was used according to the manufacturer's protocol, followed by qPCR using published primers and probes (see Methods). Western blotting for detection of phosphorylated DNA-PK, H2AX and p53 was done on total cell extracts made from uninfected or infected cells under non-denaturing conditions in the presence of phosphatase inhibitors.

Full Methods and any associated references are available in the online version of the paper.

Received 18 May 2012; accepted 9 May 2013.
Published online 5 June 2013.

1. Joint. United Nations Programme on HIV/AIDS. Global Report fact sheet: The global AIDS epidemic http://www.unaids.org/documents/20101123_FS_Global_em_en.pdf (2010).
2. Brenchley, J. M. *et al.* CD4$^+$ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J. Exp. Med.* **200,** 749–759 (2004).
3. Mattapallil, J. J. *et al.* Massive infection and loss of memory CD4$^+$ T cells in multiple tissues during acute SIV infection. *Nature* **434,** 1093–1097 (2005).
4. Nishimura, Y. *et al.* Resting naive CD4$^+$ T cells are massively infected and eliminated by X4-tropic simian-human immunodeficiency viruses in macaques. *Proc. Natl Acad. Sci. USA* **102,** 8000–8005 (2005).
5. Aiken, C. Pseudotyping human immunodeficiency virus type 1 (HIV-1) by the glycoprotein of vesicular stomatitis virus targets HIV-1 entry to an endocytic pathway and suppresses both the requirement for Nef and the sensitivity to cyclosporin A. *J. Virol.* **71,** 5871–5877 (1997).
6. Butler, S. L., Hansen, M. S. & Bushman, F. D. A quantitative assay for HIV DNA integration *in vivo. Nature Med.* **7,** 631–634 (2001).
7. Leavitt, A. D., Robles, G., Alesandro, N. & Varmus, H. E. Human immunodeficiency virus type 1 integrase mutants retain *in vitro* integrase activity yet fail to integrate viral DNA efficiently during infection. *J. Virol.* **70,** 721–728 (1996).
8. Li, L. *et al.* Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *EMBO J.* **20,** 3272–3281 (2001).
9. Chen, H., Boyle, T. J., Malim, M. H., Cullen, B. R. & Lyerly, H. K. Derivation of a biologically contained replication system for human immunodeficiency virus type 1. *Proc. Natl Acad. Sci. USA* **89,** 7678–7682 (1992).
10. Doitsh, G. *et al.* Abortive HIV infection mediates CD4 T cell depletion and inflammation in human lymphoid tissue. *Cell* **143,** 789–801 (2010).
11. Daniel, R., Katz, R. A. & Skalka, A. M. A role for DNA-PK in retroviral DNA integration. *Science* **284,** 644–647 (1999).
12. Baekelandt, V. *et al.* DNA-dependent protein kinase is not required for efficient lentivirus integration. *J. Virol.* **74,** 11278–11285 (2000).
13. Daniel, R. *et al.* Wortmannin potentiates integrase-mediated killing of lymphocytes and reduces the efficiency of stable transduction by retroviruses. *Mol. Cell. Biol.* **21,** 1164–1172 (2001).
14. Ariumi, Y., Turelli, P., Masutani, M. & Trono, D. DNA damage sensors ATM, ATR, DNA-PKcs, and PARP-1 are dispensable for human immunodeficiency virus type 1 integration. *J. Virol.* **79,** 2973–2978 (2005).
15. Callén, E. *et al.* Essential role for DNA-PKcs in DNA double-strand break repair and apoptosis in ATM-deficient lymphocytes. *Mol. Cell* **34,** 285–297 (2009).
16. Engelman, A., Mizuuchi, K. & Craigie, R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67,** 1211–1221 (1991).
17. Morozov, V. E., Falzon, M., Anderson, C. W. & Kuff, E. L. DNA-dependent protein kinase is activated by nicks and larger single-stranded gaps. *J. Biol. Chem.* **269,** 16684–16688 (1994).
18. Lau, A., Kanaar, R., Jackson, S. P. & O'Connor, M. J. Suppression of retroviral infection by the RAD52 DNA repair protein. *EMBO J.* **23,** 3421–3429 (2004).
19. Hill, R. & Lee, P. W. The DNA-dependent protein kinase (DNA-PK): More than just a case of making ends meet? *Cell Cycle* **9,** 3460–3469 (2010).
20. Nilsson, A., Sirzen, F., Lewensohn, R., Wang, N. & Skog, S. Cell cycle-dependent regulation of the DNA-dependent protein kinase. *Cell Prolif.* **32,** 239–248 (1999).
21. Chen, B. P. *et al.* Cell cycle dependence of DNA-dependent protein kinase phosphorylation in response to DNA double strand breaks. *J. Biol. Chem.* **280,** 14709–14715 (2005).
22. Nagasawa, M. *et al.* Nuclear translocation of the catalytic component of DNA-dependent protein kinase upon growth stimulation in normal human T lymphocytes. *Cell Struct. Funct.* **22,** 585–594 (1997).
23. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271,** 1582–1586 (1996).
24. Igarashi, T. *et al.* Macrophage are the principal reservoir and sustain high virus loads in rhesus macaques after the depletion of CD4$^+$ T cells by a highly pathogenic simian immunodeficiency virus/HIV type 1 chimera (SHIV): Implications for HIV-1 infections of humans. *Proc. Natl Acad. Sci. USA* **98,** 658–663 (2001).
25. Shan, L. *et al.* Stimulation of HIV-1-specific cytolytic T lymphocytes facilitates elimination of latent viral reservoir after virus reactivation. *Immunity* **36,** 491–501 (2012).
26. Ho, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373,** 123–126 (1995).
27. Wei, X. *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373,** 117–122 (1995).
28. Martínez, E. *et al.* Changes in cardiovascular biomarkers in HIV-infected patients switching from ritonavir-boosted protease inhibitors to raltegravir. *AIDS* **26,** 2315–2326 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Ambrozak, R. Nguyen, and S. Perfetto for help with cell sorting, U. Olshevsky, J. Casazza, D. Bolton, A. Pegu and M. Louder for discussions and technical help, and A. Tislerics and B. Hartman for manuscript preparation. This research was supported by the Intramural Research Program of the Vaccine Research Center, NIAID, National Institutes of Health. The findings and conclusions in this report are those of the authors and do not necessarily reflect the views of the funding agency.

Author Contributions A.C. and G.J.N. designed the research studies; A.C., M.G, C.P. and T.Y. performed the research; A.C., M.G. and T.Y. contributed to development and generation of vectors; A.C., C.P., R.A.K. and G.J.N. analysed data; and A.C., R.A.K. and G.J.N. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.J.N. (Gary.Nabel@sanofi.com).

METHODS

HIV-1 infected subjects. Samples from three HIV-1 infected, antiretroviral treatment-naive subjects with CD4 T-cell counts ranging from 24 to 609 cells per µl and viral loads ranging from 47,000 to 500,000 copies per ml were from the Vaccine Research Center study VRC200. To remove dead cells following the thawing process the samples were passed through a dead cell capture column (Pierce), resulting in >90% viability for all samples before activation. The cells were then activated by addition of PHA and IL-2 (40 U per ml) for the indicated times.

Plasmids and virus stocks. The 2 LTR circle construct used for a standard curve in the qPCR analysis was a gift from F. Bushman[6]. The HIV-$1_{89.6}$ construct was a gift from R. Collman and was previously described[29]. The HIV-1_{NL4-3} construct was obtained from M. Martin through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH[30]. The HIV-1 MC99IIIBΔTat-Rev virus was obtained from H. Chen, T. Boyle, M. Malim, B. Cullen, and H. K. Lyerly through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH[9]. The NL4-$3_{n-e-GFP}$ construct which contains a mutation in the viral envelope gene and harbours the GFP open reading frame instead of that of the viral Nef gene was a gift from M. Lenardo and was described previously[31]. NL4-$3_{n-e-GFP}$ was mutated by the Stratagene Quick Change Site-Directed Mutagenesis Kit to obtain the integrase D64V mutant used in this study. The replication-competent, GFP-encoding virus used in sorting experiments was previously described[32]. For generation of virus stocks, HIV-$1_{89.6}$ and HIV-1_{NL4-3} constructs were transfected into HEK293 T cells using the calcium phosphate method and the supernatants were collected 48 h after transfection. The viruses were then amplified for 7–10 days in CEMX174 cells and the supernatants were collected, clarified by low speed centrifugation, filtered and concentrated using Centricon Plus-70. For generation of VSVG-pseudotyped single-cycle lentiviral vectors, NL4-$3_{n-e-GFP}$ encoding either wild-type or a D64V mutant integrase were co-transfected with a plasmid encoding VSVG (pHCMV-VSVG) into 293T cells and viruses were collected and concentrated as above. Virus stocks were titred using TZM-bl cells and were frozen at −80 °C for storage.

Cell lines and primary CD4$^+$ T cell isolation. CEMX174 cells were cultured in RPMI with 10% heat-inactivated fetal bovine serum (FBS) and antibiotics (penicillin and streptomycin). HEK293T cells were cultured in DMEM supplemented with 10% FBS and antibiotics. CD4$^+$ T lymphocytes were isolated from elutriated lymphocytes prepared from blood of healthy donors by negative selection with a CD4$^+$ T cell isolation kit II (Miltenyi Biotech) according to the manufacturer's instructions. The purity of the isolated T cells was assessed by fluorescence-activated cell sorting (FACS) analysis for CD4 (BD-Pharmingen) and 95% of the cells were CD4$^+$ upon isolation. The cells were stimulated with phytohemagglutinin (0.5 µg ml^{-1}) (Remel) and 40 U ml^{-1} IL-2 (Peprotech) for 2-3 days and maintained in 20 U ml^{-1} IL-2 thereafter. CD4$^+$ T cell activation was verified using CD25 staining (BD Pharmingen) and flow cytometry.

Nucleofection and subcellular fractionation. For nucleofection of CEMX174 cells, 2 × 10^6 cells were mixed with 120 pmol of either non-targeting or human DNA ligase 4 SMARTpool siRNAs (Dharmacon) and electroporated using the Nucleofector II device (Lonza). The DNA Ligase 4 SMARTpool consisted of the following siRNA duplexes: siRNA Lig4-09 (5′-GCACAAGAUGGAGAUGUA-3′), siRNA Lig4-10 (5′-GGGAGUGUCUCAUGUAAUA-3′), siRNA Lig4-11 (5′-GGUAUGAGAUUCUUAGUAG-3′), siRNA Lig 4-12 (5′- GAAGAGGGAAUUAUGGUAA-3′). The cells were cultured for 2 days after nucleofection, and then infected. For subcellular fractionation and isolation of cell nuclei for western blot and qPCR analyses, NE-PER Nuclear and Cytoplasmic Extraction Kit (Thermo Scientific) was used according to the manufacturer's instructions.

HIV-1 infection. Purified, activated CD4$^+$ T cells were infected 2–3 days after isolation. Typically, spreading infections were done by mixing 5 × 10^5 cells with viral stocks at an m.o.i. of 0.1 followed by washing the cells 16 h after mixing. For single round infection assays, infections were typically performed by spinoculation as previously described[33]. Briefly, 0.5 × 10^6 cells were mixed with viral stock at m.o.i. = 1 in a total volume of 200 µl in a well of a flat-bottom 24-well plate and centrifuged at 1,200 × *g* for 2 h at 25 °C. Subsequently the cells were repeatedly washed and cultured at a density of 10^6 cells per ml. Mock infections were done by mixing cells with media alone. Infection of CEMX174 cells was typically done by mixing, and cells were seeded at a density of 10^6 cells per ml, except in the experiment in which antiretroviral drugs were used (2 × 10^6 cells per ml). Where indicated, the following inhibitors were used: indinavir (1 µM), efavirenz (100 nM), AZT (10 µM), raltegravir (50 µM), D-118-24 (50 µM), NU7026(20 µM) (ref. 34) , NU7441(1 µM) (ref. 35), KU55933 (10 µM), and pifithrin[36]. Cells were incubated with the inhibitors 1 h before infection or spinoculation and were present throughout the culture time. For experiments conducted for a period longer than 3 days, cell cultures were replenished with fresh drugs every 2 days.

Flow cytometric analysis of HIV-1-infected cells. CD4$^+$ T lymphocytes were harvested for Annexin V, Vivid and intracellular p24 staining at the indicated times following exposure to virus. In brief, the cells were washed and stained with Annexin V-APC (BD Pharmingen) and Vivid (Invitrogen) for 20 min. Cells were washed once, fixed and permeabilized with Cytoperm/Cytofix (BD Pharmingen) containing 2.5 ml $CaCl_2$ for 20 min. The cells were then stained for p24 Gag (FITC or KC-57 PE; Coulter) for 30 min and washed once in 1× Perm/wash buffer (BD Pharmingen). Flow cytometry was performed with an LSR II cell analyser (Becton Dickinson), and data analysis was performed with FlowJo software (Tree Star).

Quantitative viral DNA analysis. Reverse transcription products of late cellular DNA intermediates of HIV-1 reverse transcription were quantified by TaqMan real-time qPCR at the indicated times post-infection. Cellular DNA was purified using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol. Published U5/gag primers (LG564/LG699; BioSource International) and a TAMRA-labelled probe (LG-FAM; Applied Biosystems) were used to detect full-length, double-stranded viral DNA[37]. The number of copies was determined using a YU-2 plasmid as standard and as controls for linear amplification in the reaction[38]. For integrated DNA determination, a two-step PCR amplification was performed as described[39]. Briefly, a genomic Alu forward primer and an HIV-1 *gag* reverse primer were used for non-kinetic pre-amplification, followed by a second-round PCR performed using forward and reverse LTR primers and 20 µl of the material from the pre-amplification reaction. These were run with an HIV-1 copy number standard prepared from graded doses of accurately counted ACH-2 cells, which harbour a known copy number of integrated progenomes. The signal potentially contributed from unintegrated DNA was subtracted from the total signal by including a one-way preamplification with the *gag* primer alone. The 2 LTR analysis was performed using published primers and probe as previously described[6]. Primers specific for the human *β*-globin gene were used to monitor the amount of cellular DNA loaded and for normalization. Thermal cycling conditions used during the experiments were 2 min at 95° C and 40 cycles of 15 s at 95 °C and 1 min at 60 °C.

29. Collman, R. *et al.* An infectious molecular clone of an unusual macrophage-tropic and highly cytopathic strain of human immunodeficiency virus type 1. *J. Virol.* **66,** 7517–7521 (1992).
30. Adachi, A. *et al.* Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol.* **59,** 284–291 (1986).
31. Bolton, D. L. & Lenardo, M. J. Vpr cytopathicity independent of G2/M cell cycle arrest in human immunodeficiency virus type 1-infected CD4$^+$ T cells. *J. Virol.* **81,** 8878–8890 (2007).
32. Yamamoto, T. *et al.* Selective transmission of R5 HIV-1 over X4 HIV-1 at the dendritic cell-T cell infectious synapse is determined by the T cell activation state. *PLoS Pathog.* **5,** e1000279 (2009).
33. O'Doherty, U., Swiggard, W. J. & Malim, M. H. Human immunodeficiency virus type 1 spinoculation enhances infection through virus binding. *J. Virol.* **74,** 10074–10080 (2000).
34. Willmore, E. *et al.* A novel DNA-dependent protein kinase inhibitor, NU7026, potentiates the cytotoxicity of topoisomerase II poisons used in the treatment of leukemia. *Blood* **103,** 4659–4665 (2004).
35. Leahy, J. J. *et al.* Identification of a highly potent and selective DNA-dependent protein kinase (DNA-PK) inhibitor (NU7441) by screening of chromenone libraries. *Bioorg. Med. Chem. Lett.* **14,** 6083–6087 (2004).
36. Komarov, P. G. *et al.* A chemical inhibitor of p53 that protects mice from the side effects of cancer therapy. *Science* **285,** 1733–1737 (1999).
37. Suzuki, Y. *et al.* Quantitative analysis of human immunodeficiency virus type 1 DNA dynamics by real-time PCR: integration efficiency in stimulated and unstimulated peripheral blood mononuclear cells. *Virus Genes* **27,** 177–188 (2003).
38. Li, Y. *et al.* Complete nucleotide sequence, genome organization, and biological properties of human immunodeficiency virus type 1 in vivo: evidence for limited defectiveness and complementation. *J. Virol.* **66,** 6587–6600 (1992).
39. O'Doherty, U., Swiggard, W. J., Jeyakumar, D., McGain, D. & Malim, M. H. A sensitive, quantitative assay for human immunodeficiency virus type 1 integration. *J. Virol.* **76,** 10942–10950 (2002).

doi:10.1038/nature12306

cGAS produces a 2′-5′-linked cyclic dinucleotide second messenger that activates STING

Andrea Ablasser[1], Marion Goldeck[1], Taner Cavlar[1], Tobias Deimling[2], Gregor Witte[2], Ingo Röhl[3], Karl-Peter Hopfner[2,4], Janos Ludwig[1] & Veit Hornung[1]

Detection of cytoplasmic DNA represents one of the most fundamental mechanisms of the innate immune system to sense the presence of microbial pathogens[1]. Moreover, erroneous detection of endogenous DNA by the same sensing mechanisms has an important pathophysiological role in certain sterile inflammatory conditions[2,3]. The endoplasmic-reticulum-resident protein STING is critically required for the initiation of type I interferon signalling upon detection of cytosolic DNA of both exogenous and endogenous origin[4–8]. Next to its pivotal role in DNA sensing, STING also serves as a direct receptor for the detection of cyclic dinucleotides, which function as second messenger molecules in bacteria[9–13]. DNA recognition, however, is triggered in an indirect fashion that depends on a recently characterized cytoplasmic nucleotidyl transferase, termed cGAMP synthase (cGAS), which upon interaction with DNA synthesizes a dinucleotide molecule that in turn binds to and activates STING[14,15]. We here show *in vivo* and *in vitro* that the cGAS-catalysed reaction product is distinct from previously characterized cyclic dinucleotides. Using a combinatorial approach based on mass spectrometry, enzymatic digestion, NMR analysis and chemical synthesis we demonstrate that cGAS produces a cyclic GMP-AMP dinucleotide, which comprises a 2′-5′ and a 3′-5′ phosphodiester linkage >Gp(2′-5′)Ap(3′-5′)>. We found that the presence of this 2′-5′ linkage was required to exert potent activation of human STING. Moreover, we show that cGAS first catalyses the synthesis of a linear 2′-5′-linked dinucleotide, which is then subject to cGAS-dependent cyclization in a second step through a 3′-5′ phosphodiester linkage. This 13-membered ring structure defines a novel class of second messenger molecules, extending the family of 2′-5′-linked antiviral biomolecules.

Recently, it has been demonstrated that upon intracellular DNA delivery, a cytoplasmic enzyme dubbed cyclic GMP-AMP synthase (cGAS) produces a ribo-dinucleotide, which in turn binds to and activates STING[14,15]. Given the striking analogy to bacterial cyclic dinucleotide recognition and its determined molecular mass, it was suggested that this molecule constitutes a cyclic adenosine monophosphate-guanosine monophosphate (cGAMP) with a symmetric 12-membered ring formed by 3′-5′ linked nucleotide residues (>Gp(3′-5′)Ap(3′-5′)>, cGAMP(3′-5′)). On the other hand, it was shown that STING-dependent DNA sensing can be differentiated from bacterial cyclic di-GMP recognition through a point mutation at a conserved arginine residue (R231A) within the lid region of murine STING[9]. R231 functions to indirectly bind the phosphate of the phosphodiester bond of cyclic di-GMP/AMP through a Mg^{2+} or H_2O molecule, yet this coordination seems to be dispensable for STING activation in response to DNA transfection. We have recently identified a novel STING ligand (10-carboxymethyl-9-acridanone, CMA) that also triggers STING activation independently of the R231 residue[16]. In fact, the crystal structure of CMA bound to murine STING revealed that the lid region binds CMA differently than cyclic di-GMP and that R231 is not involved in CMA binding. We were intrigued by the differential role of R231 for DNA and cyclic di-GMP sensing, given the fact that modelling studies using cGAMP(3′-5′) instead of cyclic di-GMP could not readily explain the reported differential role of this residue at the structural level. To explore this further, we expressed cGAS in HEK293T cells together with either wild-type murine STING or its R231A mutant. As a control, we induced endogenous cyclic di-GMP production using a codon-optimized version of the thermophilic diguanylate cyclase domain (tDGC) (amino acids 83–248) of *Thermotoga maritima*[17] and a codon-optimized version of the recently discovered bacterial cGAMP(3′-5′) synthetase (DncV) from *Vibrio cholerae*[18]. As expected, overexpression of the cyclic di-GMP synthetase, the cGAMP synthetase and cGAS induced

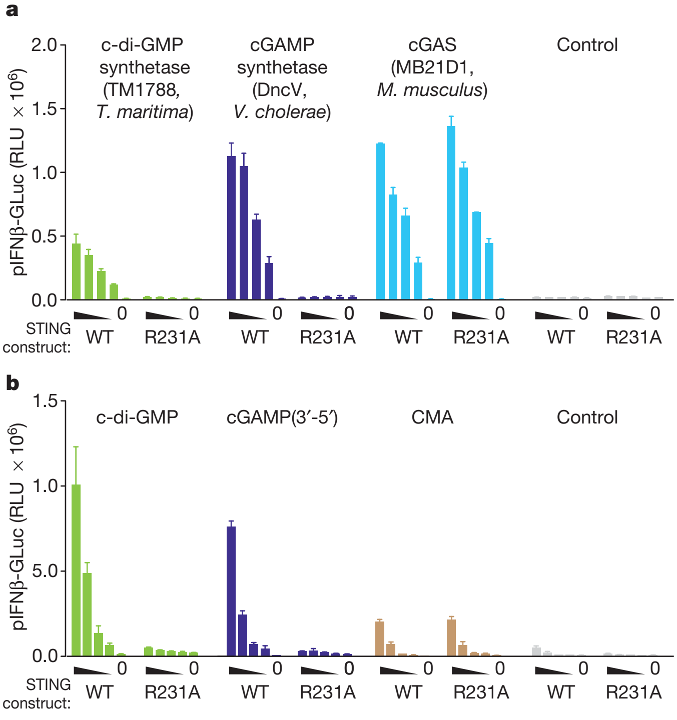


Figure 1 | The R231A STING mutant uncouples cyclic di-GMP sensing from cGAS-induced activation. a, Overexpression of dinucleotide synthetases. HEK293T cells were transfected with different dinucleotide synthetases (100 ng) together with decreasing amounts of wild-type (WT) mmSTING or the R231A mutant (10, 5, 2.5, 1.25 and 0 ng) and a pIFNβ-luciferase reporter (pIFNβ-GLuc). Reporter activity was measured 16 h after transfection. RLU, relative light units. **b**, Direct stimulation with synthetic compounds. HEK293T cells were transfected with WT mmSTING or the R231A mutant in conjunction with pIFNβ-GLuc. The next day CMA was added or synthetic cyclic di-GMP or synthetic cGAMP(3′-5′) was transfected as indicated and pIFNβ-GLuc activity was assayed 16 h later. Representative data of two (**a**) or three (**b**) independent experiments are shown (mean values + s.e.m.).

[1]Institute for Clinical Chemistry and Clinical Pharmacology, University Hospital, University of Bonn, 53127 Bonn, Germany. [2]Department of Biochemistry and Gene Center, Ludwig-Maximilians-University, 81377 Munich, Germany. [3]Axolabs GmbH, 95326 Kulmbach, Germany. [4]Center for Integrated Protein Sciences, 81377 Munich, Germany.

a robust type I interferon (IFN) response in HEK293T cells expressing wild-type murine STING. Moreover, in line with previous reports, expression of the R231A point mutant completely abolished type I IFN production in response to endogenous cyclic di-GMP production, but not upon overexpression of cGAS (Fig. 1a and Supplementary Fig. 1). Surprisingly, however, induction of endogenous cGAMP production using DncV was also completely blunted for the R231A mutant. Next we stimulated HEK293T cells overexpressing wild-type murine STING or the R231A mutant directly with synthetic compounds. As previously reported, CMA-mediated activation of STING did not require coordination through R231 and in accordance with the synthetase data from above, synthetic cyclic di-GMP only activated cells expressing wild-type murine STING, but not the R231A mutant (Fig. 1b). Unexpectedly, synthetic cGAMP(3′-5′) was also completely blunted in its stimulatory activity when transfected into cells expressing STING(R231A). Altogether, these results confirmed previous reports on DNA/cGAS-mediated STING activation being distinct from cyclic dinucleotide sensing with regards to the involvement of the lid region of STING. At the same time, however, these results questioned the concept of cGAMP(3′-5′) being the cGAS-dependent second messenger molecule activating STING.

To follow up on this observation, we generated cytoplasmic lysates from cGAS overexpressing HEK293T cells and untreated HEK293T cells and subjected the protein-depleted, low-molecular-weight fraction to reversed-phase high-performance liquid chromatography (RP-HPLC). In comparison to untreated HEK293T cells, cGAS-overexpressing HEK293T cells showed an additional, unique peak with a retention time of 46 min (Fig. 2a, *), whereas synthetic cGAMP(3′-5′) spiked into cell lysate eluted at a far higher retention time (Fig. 2a, **). Comparing endogenously produced cyclic di-GMP to synthetic cyclic di-GMP under the same conditions revealed no difference in retention time, excluding the possibility of the purification process affecting the physiochemical properties of the compounds (Supplementary Fig. 2). Fractionation of the cell-derived, cGAS-specific low-molecular-weight product (*) and transfer into STING competent LL171 cells revealed potent stimulatory activity, within the same range as synthetic cGAMP(3′-5′) (Fig. 2b and Supplementary Fig. 3). Analogous results were obtained when purified cGAS was incubated *in vitro* with GTP and ATP (Fig. 2c). A cGAS-dependent peak could be detected at the same retention time as in cell lysates from cGAS overexpressing HEK293T cells and only this peak exerted stimulatory activity in LL171 cells (Fig. 2d). Thin-layer chromatography (TLC) as an alternative separation technique revealed that the cell-derived and the *in-vitro*-synthesized cGAS product showed a similar chromatographic mobility to that of the synthetic cGAMP(3′-5′) (Fig. 2e). Despite the big difference in chromatographic properties under RP-HPLC conditions, electrospray ionization-liquid chromatography-mass spectrometry (ESI-LC-MS) analysis revealed the same molecular mass (*m*/*z* (M-H) = 673.1) for both the cell-derived cGAS product and synthetic cGAMP(3′-5′) (Fig. 2f). In addition, while the MS/MS fragmentation pattern of the cGAS-derived molecule was consistent with a ribo-dinucleotide made up of guanosine and adenosine, these studies reproducibly displayed a clear difference compared to synthetic cGAMP(3′-5′) (Supplementary Fig. 4). Most intriguingly, the MS/MS fragmentation studies pointed to the presence of a 2′-5′ phosphodiester bond between guanosine and adenosine (Supplementary Figs 4–6 and Supplementary Notes 1 and 2).

On the basis of these observations, we considered several candidate molecules as products of cGAS (Supplementary Fig. 7). Among these, a cyclic dinucleotide with one or two 2′-5′ phosphodiester bonds seemed to be most likely. To address this hypothesis, we performed a series of enzyme digests coupled to TLC and ESI-LC-MS. First, we treated synthetic cGAMP(3′-5′) and the *in-vivo*- and *in-vitro*-synthesized

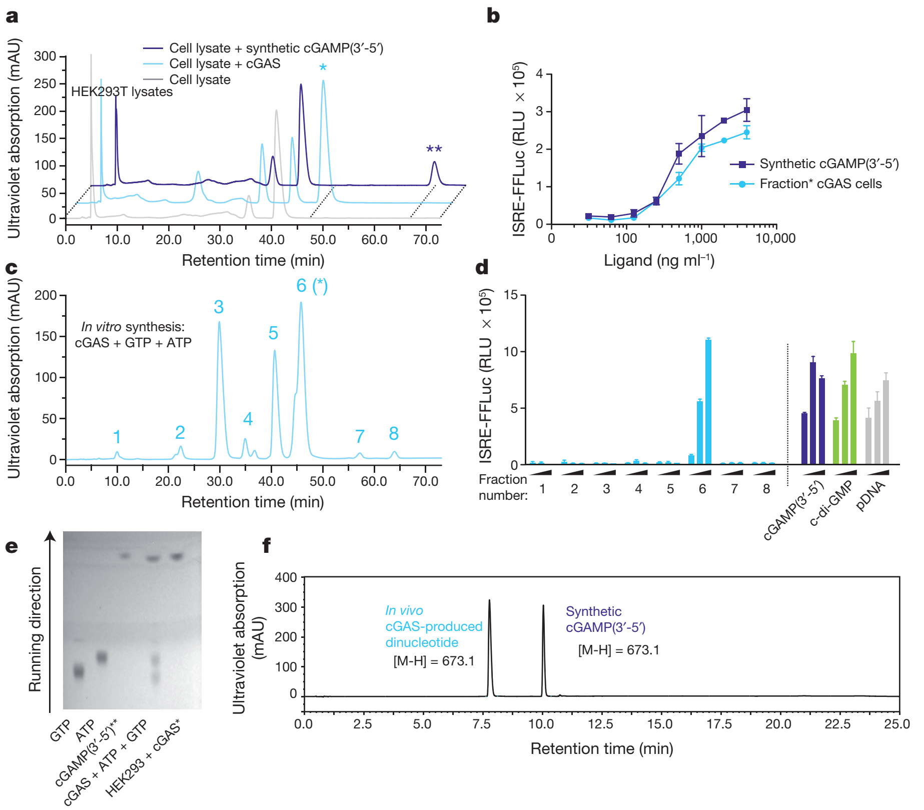


Figure 2 | The cGAS reaction product is distinct from cGAMP(3′-5′). **a**, RP-HPLC chromatograms of lysates of untreated HEK293T cells (grey), of cGAS overexpressing HEK293T cells (light blue) or of synthetic cGAMP(3′-5′) spiked into untreated HEK293T lysate (dark blue). Asterisks highlight differential elution peaks. **b**, IFN-stimulated response element (ISRE) activity in LL171 cells. Endogenous cGAS product was purified from **a** and transfected into LL171 cells, whereas synthetic cGAMP(3′-5′) served as a control. ISRE-reporter activity was measured 14 h later. **c**, Chromatogram of an *in vitro* cGAS assay. The asterisk indicates the fraction that elutes at the same retention time as the endogenous product from **a**. **d**, ISRE activity in LL171 cells. Peaks 1–8 from **c** were fractionated and transfected into LL171 cells that were then studied for ISRE-reporter activity using respective control stimuli. **e**, TLC analysis of *in-vitro-* and *in-vivo-*synthesized cGAS product with ATP, GTP and synthetic cGAMP(3′-5′) as controls. **f**, ESI-LC-MS analysis of *in-vivo*-produced cGAS product and synthetic cGAMP(3′-5′). Representative data of two (**e**) or three (**a**–**d**, **f**) independent experiments are shown (mean values + s.e.m.).

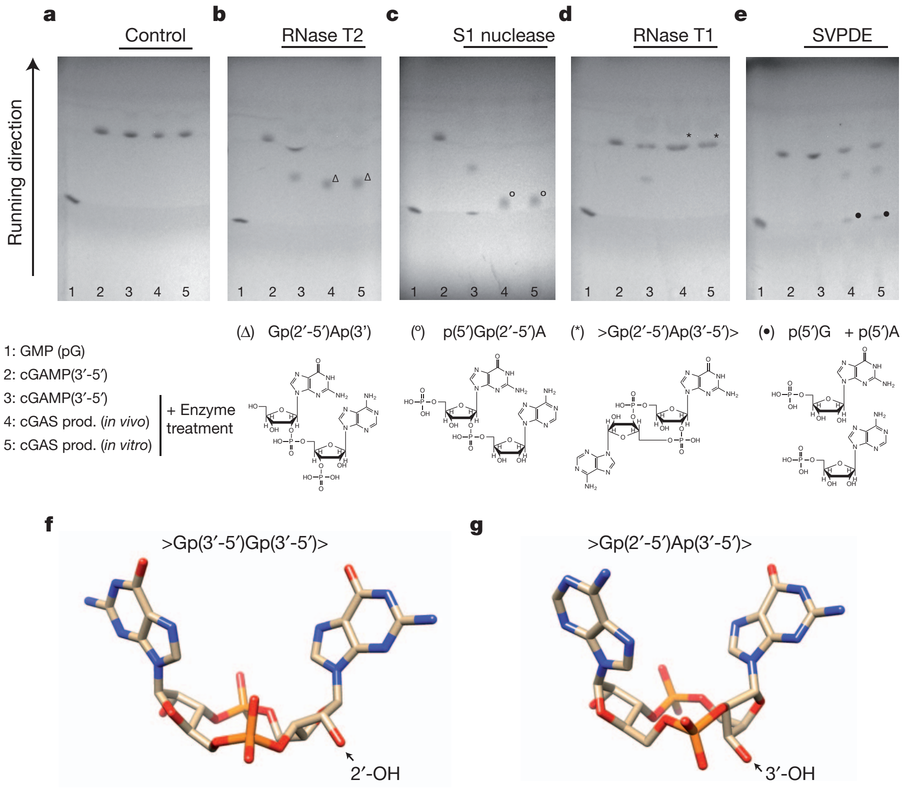


Figure 3 | The second messenger produced by cGAS is >Gp(2′-5′)Ap(3′-5′)>. **a**–**e**, TLC analysis of GMP (lane 1), synthetic cGAMP (3′-5′) (lane 2) and enzyme-treated synthetic cGAMP(3′-5′) (lane 3), *in-vivo*-synthesized cGAS product (lane 4) and *in-vitro*-synthesized cGAS product (lane 5). Enzyme treatments of molecules analysed in lanes 3–5 were control (**a**), RNase T2 (**b**), S1 nuclease (**c**), RNase T1 (**d**) and SVPDE (**e**). The resulting reaction products from lanes 4 and 5 as confirmed by ESI-LC-MS analysis are depicted below. Representative data out of two independent experiments are shown. **f**, **g**, Comparison of the structure of cyclic di-GMP (4F9G.pdb) and a model for cGAMP(2′-5′) based on NMR-derived ribose conformations.

cGAS product using S1 nuclease and ribonuclease T2. Both enzymes can cleave internal 3′-5′ phosphodiester linkages. Synthetic cGAMP(3′-5′) could be processed into mononucleotides by both enzymes, whereas the cGAS-derived cyclic dinucleotide was only cleaved into a linear dinucleotide (Fig. 3a–c and Supplementary Fig. 9a, b). These results suggested that one of the internucleotide bonds was not a 3′-5′ phosphodiester. To address which one of the two phosphodiester bonds was not hydrolysable by the enzymes above, we took advantage of the nucleotide specificity of ribonuclease T1, which catalyses the endonucleolytic cleavage of 3′-5′ phosphodiester bonds only after guanosine. As expected, ribonuclease T1 processed synthetic cGAMP(3′-5′) into a linear dinucleotide, consistent with the presence of a Gp(3′-5′)A phosphodiester bond (Fig. 3d and Supplementary Fig. 9c). The cGAS-derived dinucleotide, however, was not processed, indicating that the

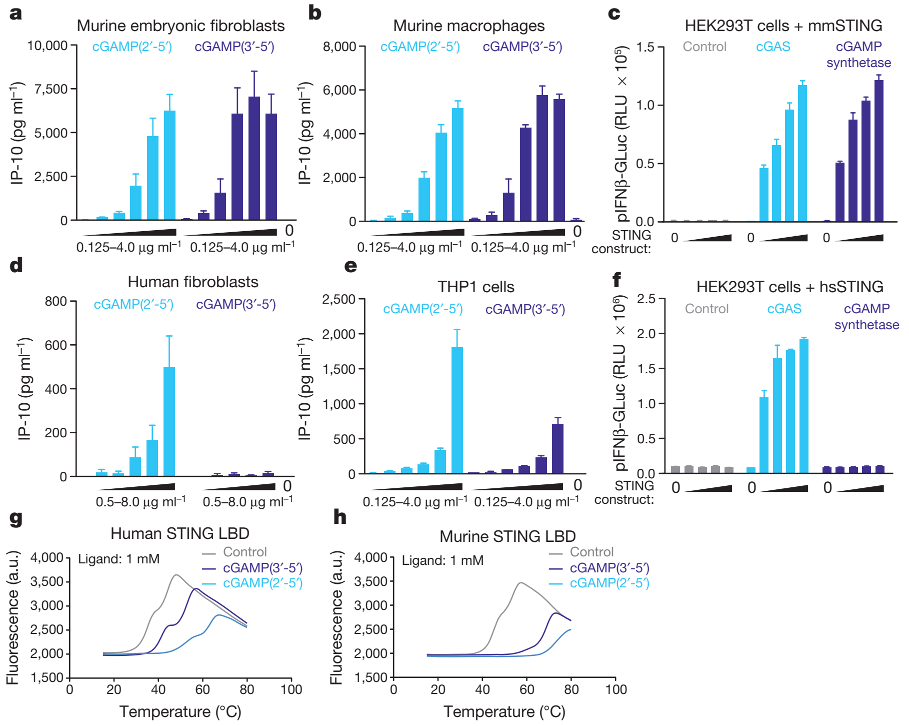


Figure 4 | cGAMP(2′-5′) is a potent activator of human and murine STING. **a**, **b**, **d**, **e**, IP-10 production of murine embryonic fibroblasts, murine macrophages, human fibroblasts and THP1 cells transfected with increasing amounts of cGAMP(2′-5′) or cGAMP(3′-5′). **c**, **f**, HEK293T cells transfected with human or murine STING (0, 3.13, 6.25, 12.5 and 25 ng) as indicated, together with cGAS or DncV cGAMP synthetase (100 ng) subsequently analysed for pIFNβ-GLuc activity. **g**, **h**, Interaction of the human and murine STING LBD with 1 mM cGAMP(2′-5′) or cGAMP(3′-5′) analysed by DSF. Mean + s.e.m. of two (**e**) or three (**a**, **b**, **d**) independent experiments or one representative out of three independent experiments (**c**, **f**, **g**, **h**) are depicted (**c**, **f**, mean values + s.e.m.).

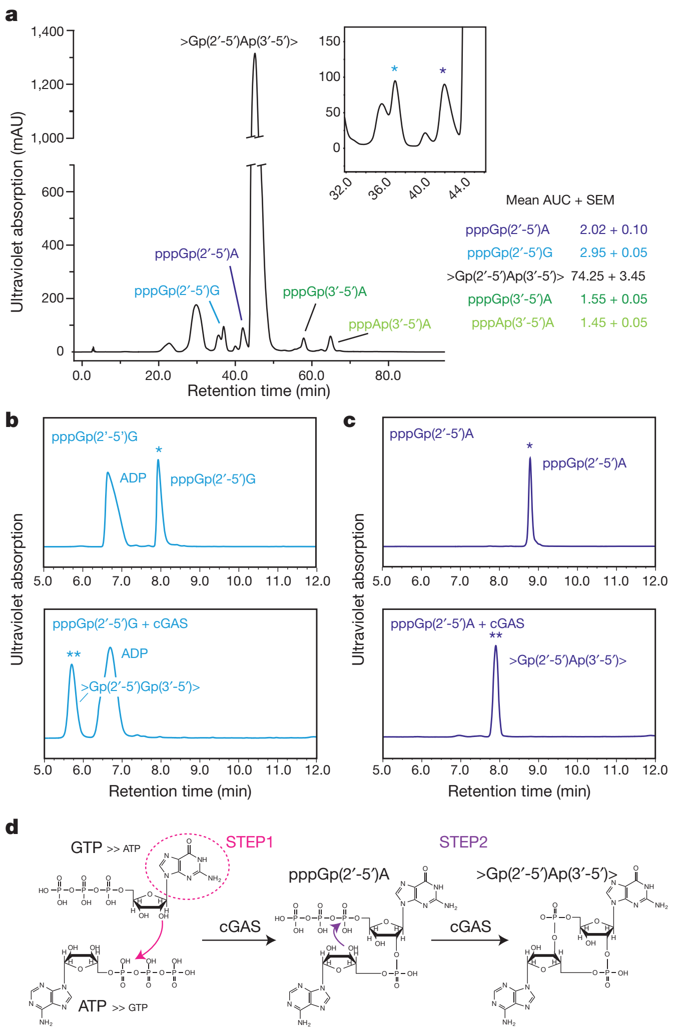


Figure 5 | >Gp(2′-5′)Ap(3′-5′)> is synthesized in a two-step process. **a**, RP-HPLC chromatogram of a cGAS+ATP+GTP *in vitro* reaction upon termination of the reaction. The insertion represents an enlargement of the chromatogram indicating the position of pppGp(2′-5′)G and pppGp(2′-5′)A. Right panel demonstrates mean area under the curve (AUC) + s.e.m. for depicted dinucleotides out of two independent experiments. **b**, **c**, ESI-LC-MS chromatograms before (top) and after (bottom) *in vitro* incubation of pppGp(2′-5′)G (**b**) and pppGp(2′-5′)A (**c**) with cGAS. Asterisks indicate the position of the substrates (*) and the resulting products (**). Note that the fraction of pppGp(2′-5′)G was contaminated with ADP. Data are representative of two independent experiments. **d**, Schematic model of the two-step process of cGAS-catalysed cyclic dinucleotide synthesis.

GpA phosphodiester bond was not 3′-5′. In line with this notion, substitution of GTP by 2′dGTP during the *in vitro* enzymatic reaction completely blunted synthesis of dinucleotides by cGAS, whereas addition of 3′dGTP gave rise to small, but consistent amounts of a cGAMP product (Supplementary Fig. 10). In a reverse approach, we made use of snake venom phosphodiesterase I (SVPDE), which can hydrolyse 5′-mononucleotides from 3′-hydroxy-terminated ribo-oligonucleotides. Consistent with its two internal 3′-5′ phosphodiester bonds, synthetic cGAMP was not processed by SVPDE, whereas the cGAS-derived product was hydrolysed in a two-step process into its mononucleotide components (Fig. 3e and Supplementary Fig. 9d). Altogether, these results clearly identified the cGAS-derived dinucleotide product as a cyclic GA dinucleotide with a 2′-5′ phosphodiester linkage between guanosine and adenosine and a 3′-5′ phosphodiester linkage between adenosine and guanosine >Gp(2′-5′)Ap(3′-5′)> (cGAMP(2′-5′)). Comparison of synthetic cGAMP(3′-5′) and the cGAS-derived product by ^{1}H-NMR spectroscopy supported the notion of differing 3′-5′/2′-5′ linkages between guanosine and adenosine within the cyclic GA dinucleotide and revealed crucial information on the ribose conformation of the distinct nucleotide elements in both molecules (Supplementary Fig. 11 and Supplementary Note 3). Relying on these data we were able to formulate a model of cGAMP(2′-5′) based on the previously determined structure of cyclic di-GMP bound to STING (Fig. 3f, g and Supplementary Note 3). To provide an additional proof of cGAMP(2′-5′) being the actual cGAS-derived second messenger molecule, we chemically synthesized >Gp(2′-5′)Ap(3′-5′)> and its isomer >Gp(3′-5′)Ap(3′-5′)> (Supplementary Fig. 12). As expected, synthetic cGAMP (2′-5′) had the same physiochemical properties as the *in vitro* cGAS-generated dinucleotide (Supplementary Fig. 13).

Transfection of both DNA and cyclic dinucleotides (3′-5′linked) has been shown to induce indistinguishable transcriptional responses in murine cells in a STING-dependent manner. However, despite being equally responsive towards DNA challenge, recent reports indicated that human cells are less responsive to intracellular delivery of cyclic dinucleotides or other STING ligands[19]. To determine whether these species-specific properties of STING would also apply for the recognition of cGAMP, we first sought to compare both cGAMP isomers with regard to their biological activity in human and murine cells. Transfection of both cGAMP(2′-5′) and cGAMP(3′-5′) into murine embryonic fibroblasts and macrophages strongly induced production of the antiviral cytokine IP-10, confirming previous reports on the stimulatory potency of cyclic dinucleotides for murine STING (Fig. 4a, b). However, to our surprise, we reproducibly observed a marked difference in responsiveness towards the two cGAMP isomers, when we tested them for functional activity in human fibroblasts and the monocytic cell line THP-1 (Fig. 4d, e). Here, cGAMP(2′-5′) was more active than cGAMP(3′-5′) with regards to production of IP-10. In fact, human fibroblasts were almost unresponsive towards transfection with cGAMP(3′-5′), even at high concentrations of the cyclic dinucleotide being delivered. Similar results were obtained when we studied HEK293T cells overexpressing cGAS or the cGAMP synthetase DncV derived from *V. cholerae* (Fig. 4c and f). Whereas expression of murine STING rendered HEK293T responsive towards both the cGAS and the cGAMP synthetase products, only cGAS expression was able to activate human STING. Finally, we performed binding studies with the carboxy-terminal ligand-binding domain (LBD) of mouse and human STING using differential scanning fluorometry (DSF). These data revealed that cGAMP(2′-5′) showed stronger complex formation with both human and mouse STING compared to cGAMP(3′-5′) (Fig. 4g, h). Of note, this preference for cGAMP(2′-5′) over cGAMP(3′-5′) was more prominent when the LBD of human STING was tested. Together, these data indicate that, in contrast to cGAMP(3′-5′), cGAMP(2′-5′) is highly potent in the human system and, like DNA, is not affected by species-specific properties of STING.

Next we wanted to delineate the mechanism of cGAS-dependent cyclic dinucleotide synthesis. Enzymatic cGAS reactions in the presence of excess substrate (ATP/GTP) displayed, among other dinucleotide species, one predominant peak in ESI-LC-MS analysis, which was represented by pppGp(2′-5′)A (Supplementary Fig. 10). However, upon termination of the reaction we identified four distinct linear dinucleotide species at almost similar quantities, next to >Gp(2′-5′)Ap(3′-5)> as the major species (Fig. 5a). These molecules included: pppGp(2′-5′)A, pppGp(3′-5′)A, pppGp(2′-5′)G and pppAp(3′-5′)A (Fig. 5a). When we incubated ATP or GTP alone with cGAS, either pppAp(3′-5′)A or pppGp(2′-5′)G predominated by approximately 10:1 over their respective phosphodiester linkage isomers (Supplementary Fig. 14). Of note, whereas GTP by itself gave rise to substantial amounts of a cyclic dinucleotide (>Gp(2′-5′)Gp(3′-5′)>), ATP by itself was unable to trigger synthesis of detectable amounts of a cyclic dinucleotide. This observation indicated that a 2′-5′ dinucleotide constitutes the substrate for the subsequent cyclization reaction and that pppGp(2′-5′)A represented the precursor molecule for cGAMP(2′-5′). To prove this hypothesis, we fractionated the four major dinucleotide species obtained during enzymatic cGAS reactions and incubated them again with cGAS. Interestingly, we found that both 2′-5′-linked dinucleotide species were quantitatively converted into cyclic dinucleotides, whereas the 3′-5′-linked dinucleotides were only scarcely, if at all, converted (Fig. 5b and Supplementary Fig. 15). Interestingly, the second phosphodiester bond was linked exclusively via 3′-5′ for all cyclization reactions. Together these results unequivocally identified pppGp(2′-5′)A as the precursor of cGAS-dependent cGAMP(2′-5′) synthesis.

On the basis of these results we postulate the following two-step synthesis model (Fig. 5d). (1) In the presence of ATP and GTP cGAS first catalyses the generation of a linear dinucleotide, with the attacking nucleotide determining the type of phosphodiester bond being generated. 5′-GTP preferentially results in a 2′-5′ linkage, whereas 5′-ATP results in a 3′-5′ linkage leading to either pppGp(2′-5′)R or pppAp(3′-5′)R, respectively. In this first synthesis step, cGAS shows a preference for GTP over ATP being the attacking nucleotide, and ATP over GTP for the nucleotide being attacked. (2) Whereas pppGp(2′-5′)R species are quantitatively cyclized by cGAS in a second step, pppRp(3′-5′)A dinucleotides are poor, if at all, substrates for cyclization. Of note, this second step exclusively generates a 3′-5′ linkage, at least for the dinucleotide species studied. The fact that only scarce amounts of cyclic di-GMP are found during *in vitro* reactions might be attributed to lower supply of its precursor molecule pppGp(2′-5′)G and presumably the preference of pppGp(2′-5′)A over pppGp(2′-5′)G during the cyclization step. All in all, this model explains the nearly exclusive generation of >Gp(2′-5′)Ap(3′-5)> by cGAS in the presence of ATP and GTP.

Previously it has been shown that the response of STING towards DNA and cyclic dinucleotides can be uncoupled[9]. This observation can now be rationalized by our finding that cGAS produces a novel class of second messenger being a 2′-5′/3′-5′-linked cyclic dinucleotide, which is structurally and physiochemically distinct from bacteria-derived cyclic dinucleotides. In fact, this report describes the enzymatic production of a cyclic 2′-5′/3′-5′-linked dinucleotide and thereby adds, at the functional level, cGAS to the oligoadenylate synthetase (OAS) family of enzymes that are unique in their ability to synthesize 2′-5′ phosphodiester bonds. Indeed, this functional similarity seems quite plausible given the sequence homology of cGAS to OAS1, which produces 2′-5′-linked oligoadenylates upon binding double stranded RNA[15,20]. Another striking analogy is that cGAS as well as the OAS enzymes both require nucleic acid binding to be activated to synthesize their products in a template-independent fashion. Thus, our results now unequivocally unify these two innate sensing systems and suggest both processes to be evolutionary linked.

The unorthodox chemical linkage within cGAMP(2′-5′) provides a unique feature that may be targeted by specific cellular regulation mechanisms. At the same time, the cGAS-dependent, two-step synthesis of cGAMP(2′-5′) could be amenable for the development of specific inhibitors for the treatment of autoimmune diseases that engage the cGAS–STING axis.

Note added in proof: After submission of the revised version of this manuscript, Gao *et al.*[21] and Diner *et al.*[22] reported the same finding, that cGAMP(2′-5′) is the cGAS-derived second messenger molecule that activates STING.

METHODS SUMMARY

Cell stimulation. If not otherwise indicated, cells were transfected using Lipofectamine 2000 (Invitrogen) with cyclic dinucleotides at a final concentration of 2 μg ml^{-1} or DNA (pCI vector) at 1.33 μg ml^{-1}.

***In vitro* assay for cGAS activity.** For *in vitro* synthesis of the cGAS reaction product 2 μM recombinant cGAS was mixed with 3 μM dsDNA in Buffer A (100 mM NaCl, 40 mM Tris pH 7.5, 10 mM $MgCl_2$) with 1 mM ATP and 1 mM GTP.

Reverse phase-HPLC. Cell lysates and enzymatic reaction mixtures were applied to a 4.1 × 250 mm PRP-1 column (Hamilton) and separated in a linear gradient of 0% buffer B for 8 min, followed by an increase of buffer B from 0 to 75% in 62 min at a flow rate of 1 ml min^{-1}. Buffer A was 20 mM triethylammonium hydrogen carbonate (TEAB) and buffer B 20 mM TEAB in 20% methanol.

Full Methods and any associated references are available in the online version of the paper.

Received 1 April; accepted 20 May 2013.
Published online 30 May 2013.

1. Hornung, V. & Latz, E. Intracellular DNA recognition. *Nature Rev. Immunol.* **10,** 123–130 (2010).
2. Gall, A. *et al.* Autoimmunity initiates in nonhematopoietic cells and progresses via lymphocytes in an interferon-dependent autoimmune disease. *Immunity* **36,** 120–131 (2012).
3. Ahn, J., Gutman, D., Saijo, S. & Barber, G. N. STING manifests self DNA-dependent inflammatory disease. *Proc. Natl Acad. Sci. USA* **109,** 19386–19391 (2012).
4. Ishikawa, H. & Barber, G. N. STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. *Nature* **455,** 674–678 (2008).
5. Zhong, B. *et al.* The adaptor protein MITA links virus-sensing receptors to IRF3 transcription factor activation. *Immunity* **29,** 538–550 (2008).
6. Jin, L. *et al.* MPYS, a novel membrane tetraspanner, is associated with major histocompatibility complex class II and mediates transduction of apoptotic signals. *Mol. Cell. Biol.* **28,** 5014–5026 (2008).
7. Sun, W. *et al.* ERIS, an endoplasmic reticulum IFN stimulator, activates innate immune signaling through dimerization. *Proc. Natl Acad. Sci. USA* **106,** 8653–8658 (2009).
8. Ishikawa, H., Ma, Z. & Barber, G. N. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* **461,** 788–792 (2009).
9. Burdette, D. L. *et al.* STING is a direct innate immune sensor of cyclic di-GMP. *Nature* **478,** 515–518 (2011).
10. Huang, Y. H., Liu, X. Y., Du, X. X., Jiang, Z. F. & Su, X. D. The structural basis for the sensing and binding of cyclic di-GMP by STING. *Nature Struct. Mol. Biol.* **19,** 728–730 (2012).
11. Ouyang, S. *et al.* Structural analysis of the STING adaptor protein reveals a hydrophobic dimer interface and mode of cyclic di-GMP binding. *Immunity* **36,** 1073–1086 (2012).
12. Shang, G. *et al.* Crystal structures of STING protein reveal basis for recognition of cyclic di-GMP. *Nature Struct. Mol. Biol.* **19,** 725–727 (2012).
13. Shu, C., Yi, G., Watts, T., Kao, C. C. & Li, P. Structure of STING bound to cyclic di-GMP reveals the mechanism of cyclic dinucleotide recognition by the immune system. *Nature Struct. Mol. Biol.* **19,** 722–724 (2012).
14. Wu, J. *et al.* Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA. *Science* **339,** 826–830 (2013).
15. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* **339,** 786–791 (2013).
16. Cavlar, T., Deimling, T., Ablasser, A., Hopfner, K. P. & Hornung, V. Species-specific detection of the antiviral small-molecule compound CMA by STING. *EMBO J.* **32,** 1440–1450 (2013).
17. Rao, F. *et al.* Enzymatic synthesis of c-di-GMP using a thermophilic diguanylate cyclase. *Anal. Biochem.* **389,** 138–142 (2009).
18. Davies, B. W., Bogard, R. W., Young, T. S. & Mekalanos, J. J. Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* **149,** 358–370 (2012).
19. Conlon, J. *et al.* Mouse, but not human STING, binds and signals in response to the vascular disrupting agent 5,6-dimethylxanthenone-4-acetic acid. *J. Immunol.* **190,** 5216–5225 (2013).
20. Kristiansen, H., Gad, H. H., Eskildsen-Larsen, S., Despres, P. & Hartmann, R. The oligoadenylate synthetase family: an ancient protein family with multiple antiviral activities. *J. Interferon Cytokine Res.* **31,** 41–47 (2011).
21. Gao, P. *et al.* Cyclic [G(2′,5′)pA(3′,5′)p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. *Cell* **153,** 1094–1107 (2013).
22. Diner, E. J. The innate immune DNA sensory cGAS produces a noncanonical cyclic dinucleotide that activates human STING. *Cell Rep.* **3,** 1355–1361 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Pelegrin for providing us with LL171 cells. K.-P.H. is supported by the National Institutes of Health (U19AI083025), the European Research Council Advanced Grant 322869, and the Center for Integrated Protein Science Munich (CIPSM). A.A. and V.H. are members of the excellence cluster ImmunoSensation. V.H. is supported by grants from the German Research Foundation (SFB670) and the European Research Council (ERC 243046).

Author Contributions A.A., M.G., T.C., G.W., T.D., I.R., J.L., K.-P.H. and V.H. designed experiments and analysed the data. A.A., M.G., T.C., G.W., T.D. and I.R. performed experiments. A.A. and V.H. wrote the manuscript. V.H. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A. (andrea.ablasser@uni-bonn.de) and V.H. (veit.hornung@uni-bonn.de).

METHODS

Reagents. Cyclic di-GMP and cyclic GAMP(3′-5′) were obtained from Biolog. DNA oligonucleotides corresponding to ISD were obtained from Metabion and annealed in PBS. 10-carboxymethyl-9-acridanone was purchased from Sigma Aldrich. ATP and GTP were obtained from Fermentas.

Cell culture. HEK293T cells, THP1 cells, human fibroblasts (hTERT-BJ1 cells), mouse embryonic fibroblasts, bone marrow-derived macrophages and LL171 cells (L929 cells containing a stable IFN-stimulated response element-luciferase reporter plasmid (ISRE-Luc)) were cultured in DMEM supplemented with 10% (v/v) FCS, sodium pyruvate (all Life Technologies) and Ciprofloxacin (Bayer Schering Pharma). All mouse cells used in this study and human hTERT-BJ1 cells and THP1 cells show responsiveness towards DNA stimulation and thus could be used for the exploration of DNA sensing pathways.

Plasmids. Expression plasmids coding for murine STING (amino-terminal green fluorescent protein (GFP)-tag)[16], murine STING R231A[16] and murine cGAS are based on pEFBOS[23]. Murine cGAS was amplified from cDNA by PCR (forward 5′-ATTACTCGAGATGGAAGATCCGCGTAGA-3′ and reverse 5′-ATTAAGATCTCTATCAAAGCTTGTCAAAAATTGGAAACCC-3′) and cloned into pEFBOS using XhoI and BglII/BamHI. A codon-optimized version of the diguanylate cyclase domain (amino acids 83–248) of TM1788 (*Thermotoga maritima* MSB8) harbouring a point mutation (R158A) to enhance c-diGMP production was cloned into pEFBOS-C-term-Flag/His using XhoI and BamHI[17]. In addition, a codon-optimized version of the *Vibrio cholerae* cGAMP synthetase (DncV; amino acids 1–438) was cloned into pEFBOS-C-term-Flag/His using XhoI and BamHI[18].

Immunoblotting. Cells were lysed in 1× Laemmli buffer and denatured at 95 °C for 5 min. Cell lysates were separated by 10% SDS–PAGE and transferred onto nitrocellulose membranes. Blots were incubated with anti-β-actin-IgG–horseradish peroxidase (HRP) and anti-GFP-IgG/anti-rabbit-IgG–HRP (all Santa Cruz Biotechnology).

Cell stimulation. LL171 cells (0.15×10^6 per ml), murine BMDM (1×10^6 per ml), MEFs (0.15×10^6 per ml), hTERT-BJ1 cells (0.2×10^6 per ml) and THP1 cells (0.6×10^6 per ml) were transfected using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Plasmid DNA (empty pCI vector) was transfected at a final concentration of 1.33 μg ml^{-1}. Unless otherwise indicated, cyclic dinucleotides were transfected at a final concentration of 2 μg ml^{-1}. Cells were stimulated 14 h before final read-out was performed.

Luciferase assay. LL171 cells were lysed in 5× passive lysis buffer (Promega) for 10 min at room temperature. The total cell lysate was incubated with firefly luciferase substrate at a 1:1 ratio and luminescence was measured on an EnVision 2104 Multilabel Reader (Perkin Elmer). pIFNβ-GLuc activity was measured in HEK293T cell supernatants using Coelenterazine as a substrate.

ELISA. Cell culture supernatants were assayed for mouse IP-10 (R&D Systems) and human IP-10 (BD Biosciences) according to the manufacturer's instructions.

***In vitro* assay for cGAS activity.** For *in vitro* synthesis of the cGAS reaction product 2 μM recombinant cGAS was mixed with 3 μM dsDNA (ISD) in Buffer A (100 mM NaCl, 40 mM Tris pH 7.5, 10 mM MgCl$_2$). Reaction was started by addition of 1 mM ATP and 1 mM GTP. After 2–4 h incubation at 37 °C the reaction was stopped and filtered using Amicon Ultra-15 filter devices (10,000 or 30,000 relative molecular mass cut-off).

Preparation of HEK293T cell lysates. HEK293T cells (0.33×10^6 per ml) were transfected with 3.2 μg plasmid using GeneJuice (Novagen). After 20 h cells were collected, washed twice with PBS and pelleted by centrifugation at 500*g* at 4 °C. The cell pellet was lysed (lysis buffer: 1 mM CaCl$_2$, 3 mM MgCl$_2$, 1 mM EDTA, 1% Triton X 100, 10 mM Tris pH 7.5) for 20 min at 4 °C. The cell lysate was briefly centrifuged (1,000*g*, 10 min, 4 °C) and the resultant supernatant was further purified via two sequential rounds of phenol-chloroform extraction. The extract was then filtered by centrifugation using Amicon Ultra-15 filter devices (10,000 or 30,000 relative molecular mass cut-off). In some experiments the final extract was concentrated via centrifugation under vacuum (Eppendorf Vacufuge).

Reversed phase-HPLC. Cell lysates and enzymatic reaction mixtures were applied to a 4.1 × 250 mm PRP-1 column (Hamilton) and separated in a linear gradient of 0% buffer B for 8 min, followed by an increase of buffer B from 0 to 75% in 62 min at a flow rate of 1 ml min^{-1}. Buffer A was 20 mM TEAB and buffer B 20 mM TEAB in 20% methanol. The product fractions were collected, evaporated and desalted by repeated co-evaporation with methanol. The residue was dissolved in PBS and the product concentration was determined by measuring ultraviolet absorbance (A_{260}). This HPLC method was mainly employed for preparative runs of cell lysates or *in vitro* synthesis products. Please note the differing retention times of this method compared to the analytical ESI-LC-MS runs.

Enzymatic reactions. 0.07 A_{260} of cGAMP(3′-5′) and cGAMP generated either *in vivo* or *in vitro* were dissolved in 6.5 μl incubation buffer and treated with 1 μl of the following enzymes: RNaseT1 (Fermentas, 100 mM Tris-HCl pH 7.4, 10 mM EDTA, 1 h, 37 °C), S1 nuclease (Fermentas, 50 mM NaOAc pH 4.5, 300 mM NaCl, 2 mM ZnSO$_4$, 1 h, 37 °C), RNase T2 (MoBiTec, 125 mM NH$_4$Ac pH 4.5, 1 h, 37 °C), SVPDE (Sigma, isolated from *Crotalus adamanteus*, 50 mM Tris-HCl pH 8.8, 10 mM MgCl$_2$, 30 min, 37 °C). The digestion products were analysed by TLC and ESI-LC-MS.

Thin layer chromatography (TLC). TLC was performed on 5 × 10 cm LuxPlate Si60 silica-covered glass plates (Merck). The samples (1–2 μl) were spotted onto the plate and separation was performed in *n*-propanol/ammonium hydroxide/water (11:7:2 v/v/v). The plate was air-dried and bands were visualized with a short-wavelength (254 nm) ultraviolet light source.

ESI-LC-MS and ESI-LC-MS/MS. All reagents used were purchased from Sigma Aldrich. The ESI-LC/MS analysis was performed using a Dionex Ultimate 3000 RS system (Thermo Fisher Scientific) coupled to an IonTrap mass spectrometer (LCQ Deca XP$^+$, Thermo Finnigan) equipped with an electrospray source operating in negative ionization mode. The ionization source parameters were set to: ion transfer capillary temperature 310 °C, spray voltage 4 kV and internal source fragmentation 15 kV. All samples were chromatographed on a Waters XBridge C18 OST column (2.1 × 50 mm; 2.5 μm particle size) at 30 °C column temperature. Separation of the analytes was achieved using a gradient of 10 mM TEAB in water as eluent A and 10 mM TEAB in 20% MeOH as eluent B with a flow rate of 0.25 ml min^{-1}. The HPLC gradient starts at 0% B, hold for 3 min and then increases over 16.5 min to 90% B.

Full-scan mass spectrometry spectra were acquired in a mass range from *m*/*z* 150 to 1,000 with isotopic resolution for the singly charged molecular ions. Tandem MS-MS and MS-MS-MS spectra were recorded from isolated ions in the ion trap applying collision induced dissociation (CID) applying helium as collision gas. For tandem MS-MS spectra the singly charged molecular ion was isolated and subsequently fragmented with 28% normalized collision energy. For tandem MS-MS-MS spectra the G-depurinated daughter ion of the cyclic dinucleotides with $m/z = 522.0$ (−guanine base) generated in the first CID fragmentation stage, was isolated and subsequently fragmented with 30% normalized collision energy.

Periodate oxidation assay. 0.1 mg ml^{-1} of the dinucleotide fractionated from cell culture lysates was incubated with 20 mM sodium periodate for 60 min at room temperature in the dark. After the incubation 10-vol% of 2 M triethylammonium acetate was added to the mixture that then was analysed by ESI-LC-MS.

2′→3′ isomerization assay. Approximately 0.1 mg ml^{-1} of the dinucleotide fractionated from cell culture lysates was incubated for 2 h at 90 °C in the presence of 10 mM EDTA and 20 mM Tris-HCl at pH 8. After the incubation 10-vol% of 2 M triethylammonium acetate was added to the mixture that then was analysed by LC-MS.

Chemical synthesis of >Gp(2′-5′)Ap(3′-5′)> and >Gp(3′-5′)Ap(3′-5′)>. The chemical synthesis of >Gp(2′-5′)Ap(3′-5′)> and >Gp(3′-5′)Ap(3′-5′)> was performed according to the strategy described in ref. 24 using the commercially available 3′-TBDMS protected 2′-guanosine phosphoramidite (Chemgenes) or 2′-TBDMS protected 3′-guanosine phosphoramidite (Sigma-Aldrich) for introduction of the 2′-5′ and 3′-5′ phosphodiester bond linkage, respectively. The 3′-adenosine phosphoramidite and all other reagents were purchased from Sigma-Aldrich. After base deprotection and removal of the TBDMS protecting groups >Gp(2′-5′)Ap(3′-5′)> or >Gp(3′-5′)Ap(3′-5′)> was purified by RP-HPLC as described above and the product was verified by ESI-LC-MS/MS.

Differential scanning fluorometry. Purification of human and murine STING ligand binding domains and differential scanning fluorometry to evaluate their thermal stabilization by nucleotide ligands was performed as previously described[16].

23. Mizushima, S. & Nagata, S. pEF-BOS, a powerful mammalian expression vector. *Nucleic Acids Res.* **18,** 5322 (1990).
24. Gaffney, B. L., Veliath, E., Zhao, J. & Jones, R. A. One-flask syntheses of c-di-GMP and the [R_p,R_p] and [R_p,S_p] thiophosphate analogues. *Org. Lett.* **12,** 3269–3271 (2010).

doi:10.1038/nature12178

Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1

Julie A. Law[1]*†, Jiamu Du[2]*, Christopher J. Hale[1]*, Suhua Feng[1,3,4], Krzysztof Krajewski[5], Ana Marie S. Palanca[6], Brian D. Strahl[5], Dinshaw J. Patel[2] & Steven E. Jacobsen[1,3,4]

DNA methylation is an epigenetic modification that has critical roles in gene silencing, development and genome integrity. In *Arabidopsis*, DNA methylation is established by DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) and targeted by 24-nucleotide small interfering RNAs (siRNAs) through a pathway termed RNA-directed DNA methylation (RdDM)[1]. This pathway requires two plant-specific RNA polymerases: Pol-IV, which functions to initiate siRNA biogenesis, and Pol-V, which functions to generate scaffold transcripts that recruit downstream RdDM factors[1,2]. To understand the mechanisms controlling Pol-IV targeting we investigated the function of SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1)[3,4], a Pol-IV-interacting protein[3]. Here we show that SHH1 acts upstream in the RdDM pathway to enable siRNA production from a large subset of the most active RdDM targets, and that SHH1 is required for Pol-IV occupancy at these same loci. We also show that the SHH1 SAWADEE domain is a novel chromatin-binding module that adopts a unique tandem Tudor-like fold and functions as a dual lysine reader, probing for both unmethylated K4 and methylated K9 modifications on the histone 3 (H3) tail. Finally, we show that key residues within both lysine-binding pockets of SHH1 are required *in vivo* to maintain siRNA and DNA methylation levels as well as Pol-IV occupancy at RdDM targets, demonstrating a central role for methylated H3K9 binding in SHH1 function and providing the first insights into the mechanism of Pol-IV targeting. Given the parallels between methylation systems in plants and mammals[1,5], a further understanding of this early targeting step may aid our ability to control the expression of endogenous and newly introduced genes, which has broad implications for agriculture and gene therapy.

SHH1 was recently identified as a Pol-IV-interacting protein and shown to affect *de novo* DNA methylation[3]. To investigate the role of SHH1 in the RdDM pathway genome-wide, we generated siRNA profiles in wild-type Col plants, *shh1* mutant plants, and several other RdDM mutants for comparison. In wild-type plants approximately 12,500 siRNA clusters were defined, representing 84.2% of all uniquely mapping 24-nucleotide siRNAs. Consistent with previous findings, 81.4% of these siRNAs were Pol-IV-dependent[6,7] (Fig. 1a; *pol-iv* and *pol-v* mutants correspond to mutations in the *nrpd1* and *nrpe1* subunits of these polymerases, respectively). Analysis of the siRNA clusters reduced in *shh1* mutants demonstrated that SHH1 is a major regulator of siRNA levels, affecting 44% of Pol-IV-dependent clusters (Fig. 1b and Supplementary Fig. 1a). These *shh1*-affected clusters represent the majority of all 24-nucleotide siRNAs, as well as a majority of clusters reduced in two downstream RdDM mutants (*drm2* and *pol-v*) (Fig. 1b and Supplementary Fig. 1a). The overlap of the reduced siRNA clusters in these mutants formed four main subclasses (termed *pol-iv* only, *shh1*, *shh1/drm2/pol-v*, and *drm2/pol-v*; Fig. 1b), which were used for subsequent analyses. Interestingly, the clusters that depend solely on Pol-IV were more enriched in pericentromeric heterochromatin than those that also depend on SHH1, DRM2 and Pol-V (Fig. 1c and Supplementary Fig. 1b, c), indicating that different mechanisms may be controlling siRNA production in the euchromatic arms versus pericentromeric heterochromatin.

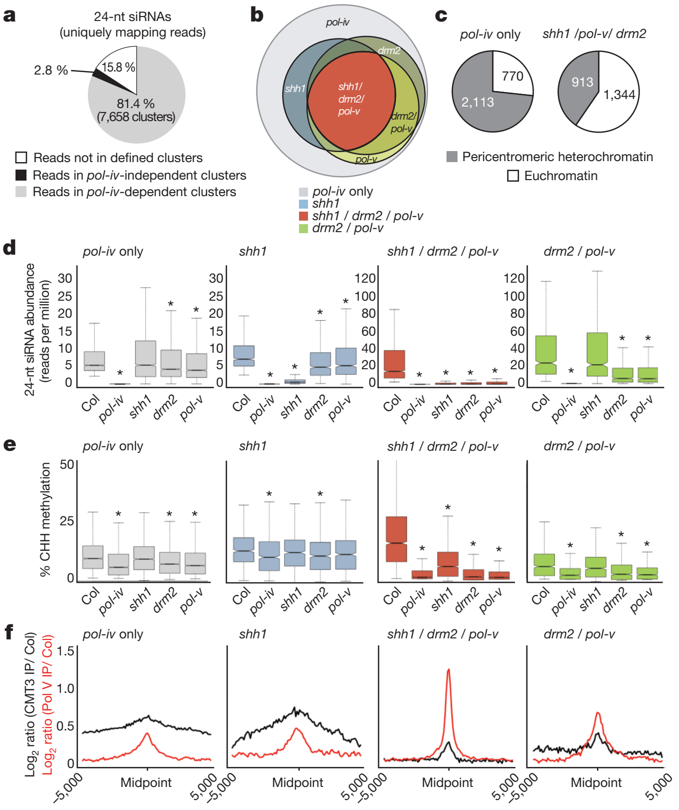


Figure 1 | Epigenetic profile of siRNA clusters affected in RdDM mutants. **a**, Pie chart showing the abundance of 24-nucleotide siRNA reads in wild-type (ecotype Col) sequencing libraries (5,967,213 uniquely mapping reads total). **b**, Schematic Venn diagram showing approximate relationships of 24-nucleotide siRNA clusters reduced in each genotype and the subclasses used for downstream analysis. **c**, Pie charts showing the chromosomal distribution (based on previously described definitions of pericentromeric heterochromatin and euchromatin[16]) of affected siRNA clusters in the indicated subclasses. **d**, **e**, Boxplots of siRNA and CHH methylation levels at the subclasses shown in **b** for various RdDM mutants (*indicates significant reduction; $P < 10^{-10}$ Mann–Whitney *U* test). **f**, Metaplots showing CMT3 and Pol-V enrichment at affected siRNA clusters (±5,000 base pairs (bp) from the siRNA cluster midpoint). IP, chromatin immunoprecipitation.

[1]Department of Molecular, Cell and Developmental Biology, University of California at Los Angeles, Los Angeles, California 90095, USA. [2]Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, California 90095, USA. [4]Howard Hughes Medical Institute, University of California at Los Angeles, Los Angeles, California 90095, USA. [5]Department of Biochemistry & Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [6]Plant Molecular and Cellular Biology, Salk Institute, La Jolla, California 92037, USA. †Present address: Plant Molecular and Cellular Biology, Salk Institute, La Jolla, California 92037, USA.

*These authors contributed equally to this work.

In *shh1* mutants, siRNA levels at SHH1-dependent clusters (*shh1* and *shh1/drm2/pol-v* subclasses) are reduced to nearly zero, whereas siRNA levels at SHH1-independent clusters experienced little to no change (Fig. 1d). These results demonstrate that SHH1 is a locus-specific RdDM component that has strong effects at a large subset of RdDM loci. Notably, the two downstream RdDM mutants (*drm2* and *pol-v*) have the strongest effect on siRNAs levels at clusters that also require SHH1 (*shh1/drm2/pol-v* subclass), and these same clusters are among the highest siRNA-producing clusters in the genome (Fig. 1d, e and Supplementary Fig. 1d, e). Together, these findings indicate that SHH1, and the downstream RdDM mutants, converge to control siRNA levels at the most active sites of RdDM.

Using whole-genome bisulphite sequencing (BS-seq), we assessed DNA methylation levels at the loci showing reduced siRNA levels and found that, consistent with its interaction with Pol-IV, SHH1 is an upstream RdDM component—*shh1* mutants only affect DNA methylation at sites where siRNA levels are reduced (Fig. 1e and Supplementary Fig. 2). Furthermore, the residual siRNAs present in *shh1* mutants seem to target some methylation (Supplementary Fig. 2b), as predicted for an upstream RdDM component. This is in contrast to the downstream mutants, *drm2* and *pol-v*, which reduced DNA methylation to nearly *pol-iv* levels even at sites that retain siRNAs (Fig. 1e), presumably due to an inability of these mutants to use siRNAs to target DNA methylation.

At loci corresponding to the *shh1/drm2/pol-v* and *drm2/pol-v* subclasses of siRNA clusters, the observed losses of siRNAs were accompanied with a correspondingly large loss of DNA methylation (Fig. 1e and Supplementary Fig. 2). However, at the *pol-iv only* and *shh1* subclasses, large losses of siRNAs were accompanied by relatively little DNA methylation loss. A likely explanation for this finding is that other DNA methylation pathways are active at sites corresponding to the *pol-iv only* and *shh1* siRNA clusters. In addition to the RdDM pathway, DNA methylation in *Arabidopsis* is controlled by two maintenance methyltransferase pathways[1]: the DNA METHYLTRANSFERASE 1 (MET1) pathway, which acts to maintain CG methylation, and the CHROMOMETHYLTRANSFERASE 3 (CMT3) pathway, which acts along with several H3K9 histone methyltransferases to maintain CHG and some CHH methylation[8]. Consistent with this explanation we found, using a previously published CMT3 chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) data set[9], that the *pol-iv* only and *shh1* subclasses of reduced siRNA clusters had the highest levels of CMT3 occupancy (Fig. 1f), indicating that CMT3 is able to maintain DNA methylation at nearly wild-type levels at these loci. In contrast, the *shh1/drm2/pol-v* and *drm2/pol-v* subclasses, which show marked DNA methylation losses in RdDM mutants, had lower levels of CMT3 enrichment (Fig. 1f) and are more highly and precisely enriched for the Pol-V polymerase[10] (Fig. 1f and Supplementary Fig. 2c), indicating that they are primarily targeted by the RdDM pathway.

To test the hypothesis that the siRNA losses observed in *shh1* mutants are due to a lack of Pol-IV targeting, we determined the genome-wide profile of Pol-IV occupancy in wild-type and *shh1* mutant backgrounds via ChIP-seq experiments using a Flag-tagged version of the largest Pol-IV subunit, NRPD1[3]. Consistent with our profile of Pol-IV-dependent siRNA clusters (Supplementary Fig. 1b), Pol-IV was broadly enriched at pericentromeric heterochromatin (Supplementary Fig. 3a) and at the defined subclasses of siRNA clusters (Fig. 2 and Supplementary Fig. 3b). In the *shh1* mutant background, Pol-IV levels were markedly reduced or eliminated specifically at *shh1*-dependent siRNA clusters (Fig. 2 and Supplementary Fig. 3c), further supporting the biological relevance of our ChIP-seq profile and confirming that the reduced-siRNA phenotype of *shh1* mutants is due to altered Pol-IV chromatin association. At *shh1*-independent siRNA clusters, Pol-IV levels, like siRNA levels, were not reduced in *shh1* mutants (Fig. 2 and Supplementary Fig. 3c), indicating that Pol-IV targeting to these loci requires an alternative mechanism.

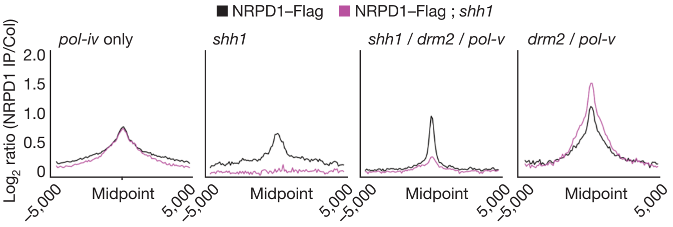


Figure 2 | Pol-IV levels at defined siRNA clusters. Metaplots of Pol-IV enrichment over the defined siRNA clusters in the indicated genetic backgrounds. Metaplots extend ±5,000 bp from the midpoint of the siRNA cluster.

In addition to assessing the levels of Pol-IV enrichment over the affected siRNA cluster subclasses, we also defined 928 reproducible, high confidence Pol-IV peaks using multiple ChIP-seq data sets. These peaks were enriched for siRNAs and DNA methylation (Supplementary Fig. 4a) and preferentially overlapped with the high siRNA-producing *shh1/drm2/pol-v* or *drm2/pol-v* clusters as compared to the *pol-iv* only and *shh1* clusters ($P < 2.2 \times 10^{-16}$, Fisher's exact test), indicating that the ChIP procedure is preferentially identifying sites where Pol-IV is most active. At the 928 defined Pol-IV peaks, we observed a variable level of SHH1 dependency and divided the peaks into three categories, SHH1-independent, SHH1-dependent and SHH1-enhanced (Supplementary Fig. 4b). In *shh1* mutants, DNA methylation and siRNA levels were reduced at the SHH1-dependent sites and, to a lesser extent, at sites defined as SHH1-independent (Supplementary Fig. 4c, d). However, siRNA and Pol-IV levels were increased at SHH1-enhanced sites in *shh1* mutants, indicating a redistribution of Pol-IV to these sites in *shh1* mutants (Supplementary Fig. 4b, c). Notably, these SHH1-enhanced sites are unique amongst the Pol-IV peaks as they have very low levels of Pol-V enrichment (Supplementary Fig. 4b), which could explain the correspondingly low levels of CHH methylation observed at these sites in wild-type plants (Supplementary Fig. 4d). Together with our analysis of SHH1-dependent siRNA clusters, these findings demonstrate that SHH1 plays a critical role in facilitating Pol-IV–chromatin association at a subset of the most active sites of RdDM.

To gain insight into the mechanism through which SHH1 facilitates Pol-IV targeting, we investigated the function of its previously uncharacterized SAWADEE domain[11]. Because there are precedents for cross talk between DNA methylation and histone modifications[1,12], we tested the ability of the SAWADEE domain to bind modified histone tails using an Active Motif-modified peptide array. This assay revealed that the SAWADEE domain has a preference for H3K9 methylation, but is also influenced by the methylation status of the H3K4 residue, with only unmodified or H3K4me1 modifications being tolerated (Supplementary Fig. 5a). To confirm these results, isothermal calorimetry (ITC) experiments were conducted using modified histone tail peptides (Fig. 3a, b and Supplementary Table 1). These analyses revealed that the SAWADEE domain is quite unique in its ability to bind all three H3K9 methylation states (me1, me2 and me3) with very similar affinity, dissociation constant (K_d) $\approx 2\,\mu$M, which is approximately 17-fold stronger than that observed using unmodified H3 peptides (Fig. 3a and Supplementary Table 1). ITC experiments also confirmed that although the SAWADEE domain will bind H3K9me2 peptides that contain H3K4me1 modifications, the presence of H3K4me2 or H3K4me3 modifications resulted in reduced binding affinity (Supplementary Table 1). Finally, ITC experiments using modified peptides corresponding to other known methylated lysine residues on the amino-terminal tails of the core histone proteins confirmed the specificity of the SHH1 SAWADEE domain for H3K9 methylation (Fig. 3b and Supplementary Table 1).

The anti-correlated effects of H3K9 and H3K4 methylation on SHH1 binding are reflective of genome profiling studies in *Arabidopsis* showing that the distribution of H3K9 methylation is anti-correlated with H3K4 methylation[13]. Consistent with these studies and the observed *in vitro*

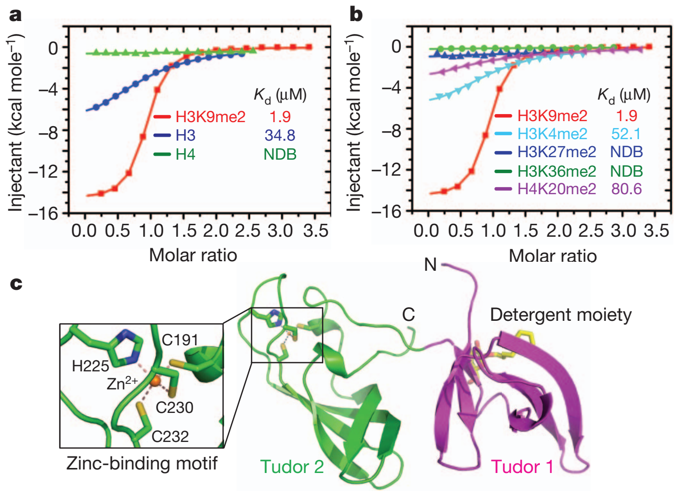


Figure 3 | The SHH1 SAWADEE domain recognizes H3K9 methylation and adopts a unique tandem Tudor domain-like fold. a, **b**, ITC-based measurements of the SAWADEE domain binding to the modified or unmodified histone peptides as indicated. K_d values are listed. NDB means no detectable binding. **c**, The overall structure of the SHH1 SAWADEE domain in the free form. The zinc-binding motif is shown as an enlarged ball-and-stick model, highlighting the details of the metal coordination. A bound detergent molecule, 4-cyclohexyl-1-butyl-β-D-maltoside moiety from the crystallization condition, is shown in a stick representation.

binding specificity of the SHH1 SAWADEE domain to H3K9 methylation, SHH1-dependent Pol-IV ChIP-seq peaks are enriched for H3K9me2 (Supplementary Fig. 5b) and depleted for H3K4 methylation (Supplementary Fig. 5c). Together, these binding studies demonstrate that the SAWADEE domain is a novel chromatin-binding module that probes both the K4 and K9 positions of the H3 tail and specifically binds repressive H3K9 methyl-modifications.

To determine the mode of methyl-lysine recognition by the SHH1 SAWADEE domain, crystal structures of this domain either in the free state or in complex with modified H3 tails were solved (Supplementary Tables 2 and 3, Fig. 3c and Supplementary Fig. 6a). In the free state, the SHH1 SAWADEE domain adopts a tandem Tudor domain-like fold that contains a unique zinc-binding motif located within the Tudor 2 subdomain (Fig. 3c). The overall structure of the SAWADEE domain resembles the UHRF1 tandem Tudor domain with an root mean squared deviation (r.m.s.d.) of 2.3 Å (Supplementary Fig. 6b) despite only sharing 11.8% sequence identity (Supplementary Fig. 7)[14,15]. This finding demonstrates that, although the sequence of the SAWADEE domain is plant-specific, its fold is highly conserved in eukaryotic organisms.

The structures of the SHH1 SAWADEE domain in complexes with H3K9me1, H3K9me2 and H3K9me3 peptides were also solved and all three peptides were bound in a similar manner (Supplementary Table 3). Given the known role of the H3K9me2 modification in gene silencing genome-wide in plants[16], we focused on the 2.70 Å structure solved with an H3(1–15)K9me2 peptide (Fig. 4a and Supplementary Fig. 8a). This peptide binds in a groove between the two Tudor subdomains, forming contacts with both subdomains (Fig. 4a, b and Supplementary Fig. 8b, c). Interestingly, there is no significant conformational change in the SAWADEE domain upon ligand binding (Supplementary Fig. 9a), which differs from the situation for UHRF1 (ref. 15).

Within the SHH1 SAWADEE domain, there are two pockets that form key intermolecular interactions with the unmodified K4 and the K9me2 side chains of the bound peptide (Fig. 4c, d). The unmodified H3K4 side chain inserts into an interfacial pocket formed by residues from both Tudor subdomains. In this pocket, the K4 side chain is stabilized via intermolecular hydrogen bonds and electrostatic interactions with the side chains of Glu 130 and Asp 141 (Fig. 4c). The H3K9me2 side chain inserts into a hydrophobic aromatic cage in the Tudor 1 subdomain (Fig. 4d) where it is stabilized by cation-π interactions in a manner similar to those reported previously for methylated lysine-binding modules[17]. The SAWADEE complexes with H3K9me3 and H3K9me1 peptides also position the methylated lysines within the same aromatic cage (Supplementary Fig. 10). The ability of the SAWADEE domain to bind equally against all three H3K9 methylation states can be well explained by structural observations: The methylated lysine recognition aromatic cage can accommodate both H3K9me2 and H3K9me3 side chains through common hydrophobic interactions, resulting in a lack of discrimination between these two methylation states. In the H3K9me1 complex, although the lower lysine methylation state has a decreased hydrophobic interaction with the aromatic cage, the side chain of His 169 undergoes a small but significant conformational change in order to hydrogen bond with the K9me1 ammonium proton, thereby contributing to the recovery of the binding affinity (Supplementary Fig. 10). This lack of specificity for the state of K9 methylation is in contrast with the higher level of methylation specificity observed for the tandem Tudor domain of UHRF1, which has a slightly wider aromatic cage binding pocket (Supplementary Fig. 9b). Thus our structural analysis indicates how very subtle changes in the tandem Tudor domain fold can result in a fine tuning of methyl-lysine specificity.

Consistent with our peptide-binding studies (Supplementary Table 1), we were also able to solve a structure of the SAWADEE domain in a complex with an H3(1–15)K4me1K9me1 peptide (Supplementary Table 3). Overall, this structure resembles the structure with the H3K9me2 peptide, with the K4me1 accommodated within the same K4 binding pocket. However, the methyl group forms a stabilizing hydrophobic interaction with Leu 201 in place of the hydrogen bond that is formed between the unmethylated K4 and the Glu 130 side chain (Fig. 4e). Because this K4 binding pocket is relatively closed and narrow, higher methylation states of K4 would probably introduce steric conflicts and/or disrupt all the hydrogen bonding interactions, explaining the observed decreases in binding affinity (Supplementary Table 1).

To test the biological significance of methyl-H3K9 binding activity observed for the SHH1 SAWADEE domain, we generated point mutations within the two lysine-binding pockets as well as the zinc-binding motif and tested their effect on DNA methylation, siRNA levels and Pol-IV recruitment *in vivo*. These point mutations were engineered into an SHH1–3×Myc–BLRP (biotin ligase recognition peptide) construct and transformed into an *shh1* mutant background. DNA methylation levels were assessed at a well-characterized locus, *MEA-ISR*, by Southern blotting (Supplementary Fig. 11a) and genome-wide by BS-seq experiments (Fig. 4f and Supplementary Fig. 11c–e). Addition of a wild-type SHH1–3×Myc–BLRP transgene restored DNA methylation, but constructs harbouring mutations within the H3K9 or the H3K4 pockets were unable to fully complement the methylation defect observed in the *shh1* mutant (Fig. 4f and Supplementary Fig. 11c–e) despite being expressed at levels comparable to the wild-type SHH1–3×Myc–BLRP protein (Supplementary Fig. 11a). In line with a canonical role for the zinc-binding motif in protein structure and/or stability, mutations in the zinc coordinating residues resulted in nearly undetectable levels of protein (Supplementary Fig. 11b) and thus were not characterized further.

Similar to the *shh1* null mutant, the DNA methylation defects in the SHH1 lysine binding pocket mutants were most pronounced in the *shh1/drm2/pol-v* subclass of affected siRNA clusters (Fig. 4f and Supplementary Fig. 11c–e). Consistent with their positions and predicted contributions to the binding affinity of the SHH1 SAWADEE domain, the F162AF165A and the D141A mutants show stronger DNA methylation defects (Fig. 4f). Assessment of siRNA levels in these lysine binding pocket mutants via siRNA-seq experiments revealed a similar pattern of defects (Fig. 4g and Supplementary Fig. 11f). Finally, to determine whether the observed losses of siRNAs and DNA methylation reflect a defect in Pol-IV activity at chromatin, Pol-IV ChIP experiments were conducted in the SAWADEE domain point mutant backgrounds. All

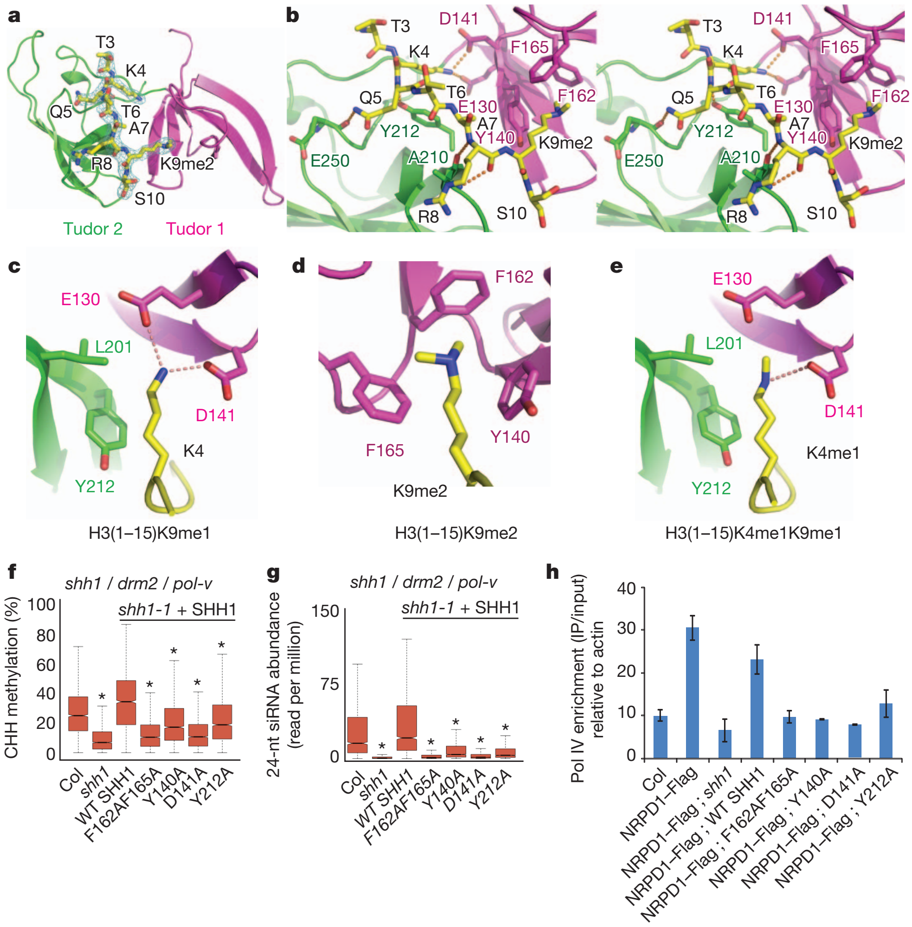


Figure 4 | Structural basis for recognition of H3(1–15)K9me2 peptide by the SHH1 SAWADEE domain and the functional impact of mutations of residues lining the K4 and K9me2 pockets. **a**, Overall structure of the H3(1–15)K9me2–SAWADEE complex with the SAWADEE domain as a ribbon diagram and the peptide as a stick representation . The simulated annealing composite omit map at 1*σ* level of the bound peptide is also shown. **b**, Stereo view highlighting the intermolecular interactions between the SAWADEE domain and the bound peptide. Intermolecular hydrogen-bonding interactions are designated by dashed red lines. **c**–**e**, Close-up views of H3 lysine residues (**c**, H3(1–15)K9me1; **d**, H3(1–15)K9me2; **e**, H3(1–15)K4me1K9me1) in their respective binding pockets. **f**, **g**, Boxplots of genome-wide percentage CHH methylation and siRNA levels in wild-type, *shh1* mutants and *shh1* mutants transformed with *SHH1* constructs (*shh1* + SHH1) that encode wild-type SHH1 or K9 (F162AF165A and Y140A) or K4 (D141A and Y212A) binding pocket mutants. **h**, qPCR of Pol-IV enrichment in the backgrounds described in **f** at a defined Pol-IV binding site. Bars are the average of two biological replicates normalized to input and actin levels (± standard error).

four point mutants displayed reduced levels of Pol-IV occupancy in two biological replicates (Fig. 4h). In addition, co-immunoprecipitation experiments revealed that the SAWADEE domain point mutants were still able to interact with Pol-IV (Supplementary Fig. 11g), demonstrating the interaction between SHH1 with the Pol-IV complex is not dependent on its H3K9me binding activity. Together, these findings show that residues within both the K4 and K9 binding pockets are critical for SHH1 function *in vivo* and demonstrate a central role for methyl-H3K9 binding by SHH1 at the level of Pol-IV association with chromatin.

The finding that the H3K4 binding pocket is critical for SHH1 function *in vivo* was unexpected considering that the SHH1 SAWADEE does not bind H3K4 methylation in the absence of H3K9 methylation, and that the addition of a methyl group to K4 does not impart additional binding affinity (Supplementary Table 1). One hypothesis to explain these *in vivo* findings is that the mere presence of a lysine at the position five residues back from the methylated H3K9 residue is necessary for SAWADEE domain binding. Indeed, such dual lysine reading could serve to help ensure that the SAWADEE domain only binds lysine methylation when it is present at the K9 position of the H3 tail as opposed to a methylated lysine at a different position on the H3 tail, especially the H3K27 position which has similar ARKS sequence context as H3K9 but a Thr 22 at five residues back. To test this hypothesis, ITC experiments were conducted using H3 tails harbouring an H3K4A mutation with or without the presence of the H3K9me2 modification. Indeed, the SAWADEE domain binds the H3K4AK9me2 peptide with approximately 30-fold weaker affinity than the H3K9me2 peptide (Supplementary Table 1). Furthermore, the SHH1 SAWADEE domain binds the H3K4A peptide with weaker affinity than the wild-type H3 tail (Supplementary Table 1), demonstrating that the K4 residue is contributing to binding independent of the methylation status of the K9 residue.

Together, these *in vivo* and *in vitro* analyses demonstrate that the SHH1 SAWADEE domain is probing the H3 tail at both the K4 and K9 positions and is quite selective for the combination of histone modifications present at transposons and other repetitive DNA elements, namely unmodified H3K4 and methylated H3K9. Although H3K9 methylation is anti-correlated with H3K4 methylation genome-wide[13], the aversion of the SAWADEE domain to higher order H3K4 methylation could serve to allow transcription, which is correlated with H3K4 methylation, to overcome DNA methylation and associated repressive H3K9 methyl modifications in a developmental or locus-specific manner. Likewise, the specificity of the SAWADEE domain could inhibit siRNA generation at body methylated genes which contain CG methylation and H3K4 methyl-modifications, but lack CHG and CHH methylation as well as siRNAs[13,18,19].

In summary, we demonstrate that SHH1 is a novel chromatin-binding protein that functions to enable Pol-IV recruitment and/or stability at the most actively targeted genomic loci to promote siRNA biogenesis. The finding that SHH1 binds to repressive histone modifications, together with the observation that SHH1 is required for Pol-IV chromatin association at a similar set of loci as downstream RdDM mutants, could explain the previously observed self-reinforcing loop in which downstream RdDM mutants are required for the production of full levels of siRNAs from a subset of genomic loci[20–23]. Indeed, it has been shown that downstream RdDM mutants can cause a reduction of both DNA methylation and H3K9 methylation at RdDM loci[24], suggesting that the loss of siRNAs in these mutants may be due to the associated loss of the appropriate chromatin marks necessary for SHH1 binding.

METHODS SUMMARY

Materials and methods for histone peptide array, ITC binding, crystallization, structure determination and analysis, plant lines, and genomic data analysis are described in detail in the Methods. Read statistics for all genomics analyses are listed in Supplementary Table 4, and the defined small RNA clusters and Pol-IV peaks are listed in Supplementary Tables 5 and 6, respectively.

Full Methods and any associated references are available in the online version of the paper.

Received 2 September 2012; accepted 12 April 2013.
Published online 1 May 2013.

1. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11,** 204–220 (2010).
2. Haag, J. R. & Pikaard, C. S. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nature Rev. Mol. Cell Biol.* **12,** 483–492 (2011).
3. Law, J. A., Vashisht, A. A., Wohlschlegel, J. A. & Jacobsen, S. E. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genet.* **7,** e1002195 (2011).
4. Liu, J. *et al.* An atypical component of RNA-directed DNA methylation machinery has both DNA methylation-dependent and -independent roles in locus-specific transcriptional gene silencing. *Cell Res.* **21,** 1691–1700 (2011).
5. Olovnikov, I., Aravin, A. A. & Fejes Toth, K. Small RNA in the nucleus: the RNA-chromatin ping-pong. *Curr. Opin. Genet. Dev.* **22,** 164–171 (2012).
6. Mosher, R. A., Schwach, F., Studholme, D. & Baulcombe, D. C. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl Acad. Sci. USA* **105,** 3145–3150 (2008).
7. Zhang, X., Henderson, I. R., Lu, C., Green, P. J. & Jacobsen, S. E. Role of RNA polymerase IV in plant small RNA metabolism. *Proc. Natl Acad. Sci. USA* **104,** 4536–4541 (2007).
8. Cao, X. *et al.* Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Curr. Biol.* **13,** 2212–2217 (2003).
9. Du, J. *et al.* Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151,** 167–180 (2012).
10. Zhong, X. *et al.* DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nature Struct. Mol. Biol.* 19**,** 870–875 CrossRef (2012).
11. Mukherjee, K., Brocchieri, L. & Burglin, T. R. A comprehensive classification and evolutionary analysis of plant homeobox genes. *Mol. Biol. Evol.* **26,** 2775–2794 (2009).
12. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10,** 295–304 (2009).
13. Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M. & Jacobsen, S. E. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis* thaliana. *Genome Biol.* **10,** R62 (2009).
14. Holm, L. & Rosenstrom, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38,** W545–W549 (2010).
15. Nady, N. *et al.* Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. *J. Biol. Chem.* **286,** 24300–24311 (2011).
16. Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M. & Jacobsen, S. E. Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. *PLoS ONE* **3,** e3156 (2008).
17. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Struct. Mol. Biol.* **14,** 1025–1040 (2007).
18. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis. Cell* **126,** 1189–1201 (2006).
19. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
20. Zilberman, D. *et al.* Role of *Arabidopsis* ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr. Biol.* **14,** 1214–1220 (2004).
21. Xie, Z. *et al.* Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2,** e104 (2004).
22. Li, C. F. *et al.* An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* **126,** 93–106 (2006).
23. Pontes, O. *et al.* The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126,** 79–92 (2006).
24. Zilberman, D., Cao, X. & Jacobsen, S. E. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299,** 716–719 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the staff of beamlines 24ID-C/E at the Argonne National Laboratory and X29A at the Brookhaven National Laboratory for support in diffraction data collection, to M. Akhavan and S. Fuchs for technical assistance, and to the UCLA BSCRC BioSequencing Core Facility for DNA sequencing. This work was supported by funds from the Abby Rockefeller Mauze Trust and Maloris and STARR Foundations to D.J.P. Work in the Jacobsen laboratory was supported by NIH grant GM60398. C.J.H. is supported by the Damon Runyon postdoctoral fellowship. S.F. is a Special Fellow of the Leukemia & Lymphoma Society. B.D.S. is supported by NIH grant GM85394. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions J.A.L., J.D., C.J.H., S.F. and A.M.S.P. conducted the experiments, K.K. synthesized modified peptides, B.D.S., D.J.P. and S.E.J. directed the research, and J.A.L., J.D., C.J. H., D.J.P. and S.E.J. wrote the manuscript.

Author Information Coordinates and structure factors have been deposited in the RCSB Protein Data Bank with the accession codes: 4IUP for the Se–SAWADEE (L200M, L218M) complex and 4IUQ for the wild-type SAWADEE domain in the free state, 4IUR for the H3(1–15)K9me3–SAWADEE complex, 4IUT for the H3(1–15)K9me2–SAWADEE complex, 4IUU for the H3(1–15)K9me1–SAWADEE complex, and 4IUV for the H3(1–15)K4me1K9me1–SAWADEE complex. The genomics data are submitted to GEO under the accession number GSE45368. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.E.J. (jacobsen@ucla.edu) or D.J.P. (pateld@mskcc.org).

METHODS

ChIP-seq, BS-seq and siRNA-seq library construction and sequencing. The first replicate of ChIP-seq libraries (NRPD1–Flag and Col) was generated using the Ovation Ultralow IL Multiplex System (NuGEN) whereas the second replicate (NRPD1–Flag, NRPD1–Flag ; *shh1*, and Col) was generated using the Ovation Ultralow DR Multiplex System (NuGEN). Both sets of ChIP-seq libraries used 18 cycles for the library amplification step. BS-seq libraries were generated using the pre-methylated adapter method as described previously[25]. siRNA-seq libraries were generated using the small RNA TruSeq kit (Illumina) following the manufacturer instructions with the exception that 15 cycles were used during the amplification step. The wild-type (Col) and *nrpe1* BS-seq libraries used in this study were previously published[10] and were subsequently reanalyzed for this study. All libraries were sequenced using the HiSeq 2000 platform following manufacturer instructions (Illumina) at a length of 50 bp. Read statistics are listed in Supplementary Table 4.

Mapping and processing of reads. Sequenced reads were base-called using the standard Illumina pipeline. For ChIP-seq and BS-seq libraries, only full 50 nucleotides reads were retained, whereas for siRNA-seq libraries, reads had adapter sequence trimmed and were retained if they were between 18 and 28 nucleotides in length. For ChIP-seq and siRNA-seq libraries, reads were mapped to the *Arabidopsis* genome (TAIR8, http://www.arabidopsis.org) with Bowtie[26] and only perfect matches that mapped uniquely to the genome were retained for further analysis although the total number of mapping reads, unique and non-unique, were used when normalizing the siRNA-seq libraries to total number of reads per library. For BS-seq libraries, reads were mapped using the BSseeker wrapper for Bowtie[27]. For ChIP-seq and BS-seq, identical reads were collapsed into one read, whereas for siRNA-seq identical reads were retained. For methylation analysis, percent methylation was calculated as previously reported[19] with the unmethylated chloroplast genome serving as the measure of non-bisulphite converted background methylation. For the second replicate of ChIP-seq, there was a large disparity of resultant reads for the NRPD1–Flag and NRPD1–Flag ; *shh1* libraries, so the NRPD1–Flag and Col libraries were sampled down to match the read total of the smaller library (the NRPD1–Flag ; *shh1* library).

DNA methylation analysis. For assessment of DNA methylation at siRNA clusters, only those clusters with at least one cytosine in the respective class being assayed (CG, CHG or CHH), were considered. For calculating significance levels of methylation change via the Mann–Whitney *U* test of methylation levels for clusters within the different subclasses (Fig. 1e) the number of clusters within each subclass was down sampled to the smallest subclass (the *drm2/nrpe1* subclass) to allow for comparable significance values between subclasses.

Identification of siRNA clusters. Small RNA clusters (Supplementary Table 5) in the *Arabidopsis* genome were defined in a manner similar to a previously published approach[28]. In brief, the genome was divided into 200-bp bins, and the average coverage per bin of non-identical siRNA reads was calculated in two technical replicates of our wild-type (Col) library. This average was used to assay the significance of the number of non-identical reads at a given bin in wild-type plants, assuming a Poisson distribution of such counts. In the R environment a Poisson exact test was carried out for each bin, and bins with a *P*-value less than 10^{-5} in each wild-type technical replicate were considered as clusters.

Once clusters were defined, comparisons between read counts, including identical reads, were carried out for each mutant and the wild-type (Col) library using a Fisher's exact test. Resultant *P*-values were Benjamini–Hochberg adjusted to estimate false discovery rates (FDRs), and clusters reduced in a mutant background at a FDR $< 10^{-10}$ were then considered to be dependent on the wild-type function of the mutant protein (Supplementary Table 5). For boxplot analysis of siRNA levels, the first technical replicate of the Col library was used as representative of Col siRNA levels. For calculating significance levels of siRNA change via the Mann–Whitney *U* test of siRNA levels for clusters within the different genotypic subclasses (Fig. 1d) the number of clusters within each subclass was down sampled to the smallest subclass (the *drm2/nrpe1* subclass) to allow for comparable significance values between subclasses.

Identification of NRPD1 peaks. The R package BayesPeak[29,30] was used to identify regions of Pol-IV enrichment in a NRPD1–Flag ChIP-seq library as compared to a paired Col ChIP-seq control library done in parallel. Only high scoring peaks (PP > 0.999) identified in both NRPD1–Flag ChIP-seq replicates (928 peaks) were retained for further analysis (Supplementary Table 6). For the purposes of assaying overlap of Pol-IV peaks with siRNA clusters, 'overlap' is called when more than 1 bp of a peak overlaps with a locus.

To classify peaks as SHH1-dependent, -independent or -enhanced, read counts over Pol-IV peaks were compared between the NRPD1–Flag and NRPD1–Flag; *shh1* ChIP-seq libraries, and significance was assessed using Fisher's exact test. Resultant P values were Benjamini–Hochberg adjusted to estimate FDRs. Peaks with a loss of NRPD1 signal in the *shh1* library at a FDR < 0.001 were considered SHH1-dependent. Similarly, peaks that gained signal in *shh1* at a FDR < 0.001 were considered SHH1-enhanced. Peaks that fell into neither of these categories were considered SHH1-independent.

Protein preparation. The gene encoding the SAWADEE domain of *Arabidopsis thaliana* SHH1 (residues 125–258) was cloned into a self-modified vector, which fuses a hexa-histidine tag plus a yeast sumo tag onto the N terminus of the target gene. The plasmid was transformed into the *Escherichia coli* strain BL21 (DE3) RIL (Stratagene). The cells were grown at 37 °C until the $D_{600\,\mathrm{nm}}$ reached 0.8 and then the media was cooled to 20 °C and 0.2 mM IPTG was added to induce protein expression overnight. The recombinant expressed protein was first purified using a HisTrap FF column (GE Healthcare). The hexa-histidine-sumo tag was cleaved by the Ulp1 protease and removed by passing through a second HisTrap FF column. The pooled target protein was further purified using a Q FastFlow column and a Hiload Superdex G200 16/60 column (GE Healthcare) with buffer (150 mM NaCl, 20 mM Tris pH 8.0, and 5 mM dithiothreitol (DTT)). To prepare the Se-methionine-substituted protein, Leu 200 and Leu 218 of the SAWADEE domain were mutated to methionine using a QuikChange Site Directed Mutagenesis kit (Stratagene). The Se-methionine-substituted SAWADEE protein was expressed in M9 medium supplemented with amino acids Lys, Thr, Phe, Leu, Ile, Val and Se-Met, and purified using the same protocol as the wild-type protein. Peptides were synthesized by the Tufts University peptide synthesis facility or by K. Krajewski.

Crystallization. Crystallization of the SAWADEE domain was conducted at 4 °C using the sitting drop vapour diffusion method by mixing 1 μl of protein sample at a concentration of 5 mg ml^{-1} and 1 μl of reservoir solution (0.2 M NH_4F and 20% PEG 3350), which was equilibrated against a 0.4 ml reservoir. 4-cyclohexyl-1-butyl-β-D-maltoside (CYMAL-4, Hampton Research) was added in the drop with a final concentration of 7.6 mM as an additive, which resulted in considerable improvement in crystal quality. Thin plate-shaped crystals appeared within 2 days. To generate crystals of complexes of SAWADEE domain with modified H3 peptides (H3(1–15)K9me3, H3(1–15)K9me2, H3(1–15)K9me1 and H3(1–15)K4me1K9me1), the SAWADEE domain was mixed with peptides at a molar ratio of 1:2 at 4 °C for 1 h. The crystals of the different complexes were grown under the same conditions as described for free SAWADEE protein. All the crystals were soaked into a reservoir solution supplemented with 20% glycerol for 2 min. The crystals were then mounted on a nylon loop for diffraction data collection. The diffraction data from the native SAWADEE protein and its Se-methionine-substituted counterpart were collected at the NE-CAT beamline 24ID-C, Advanced Photon Source (APS), Argonne National Laboratory, Chicago, USA, at the zinc peak and selenium peak, respectively. The data of the complex of the H3K9me3 peptide bound to the SAWADEE domain were collected at beamline X29A, National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory, New York, USA. The data on the SAWADEE domain in complex with H3K9me2, H3K9me1 and H3K4me1K9me1 peptides were collected at APS 24ID-E. All the crystallographic data were processed with the HKL2000 program[31]. The statistics of the diffraction data are summarized in Supplementary Tables 2 and 3.

Structure determination and refinement. The structure of the selenomethionine-substituted SAWADEE domain was solved using the single-wavelength anomalous dispersion (SAD) method as implemented in the Phenix program[32]. The model building was carried out using the Coot program[33] and structural refinement using the Phenix program[32]. The structure of the wild type SAWADEE domain in the free state was solved using the molecular replacement method using the Phenix program[32]. Zn^{2+} ions were identified and further confirmed by anomalous signal scattering. All the structures of SAWADEE domain in complexes with different modified H3 peptides were solved using the molecular replacement method with the same protocol as the native protein. The peptides showed clear electron density and were properly built with residues from Thr 3 to Ser 10 for H3(1–15)K9me3/2/1 and from Thr 3 to Thr 11 for H3(1–15)K4me1K9me1. Throughout the refinement, a free *R* factor was calculated using 5% random chosen reflections. The stereochemistry of the structural models were analysed using the Procheck program[34]. The refinement and structure statistics are shown in Supplementary Tables 2 and 3. All the molecular graphics were generated with the Pymol program (DeLano Scientific LLC).

Isothermal titration calorimetry. The protein samples were not stable at room temperature. Thus, all the binding experiments were performed on a Microcal calorimeter ITC 200 instrument at 6 °C. First, protein samples were dialysed overnight against a buffer of 100 mM NaCl, 2 mM β-mercaptoethanol and 20 mM HEPES, pH 7.5, at 4 °C. Then the protein samples were diluted and the lyophilized peptides were dissolved with the same buffer. The titration was performed according to standard protocol and the data were fit using the Origin 7.0 program with a 1:1 binding model. Thermodynamic parameters for complex formation are listed in Supplementary Table 1.

Modified peptide array binding. A glutathione *S*-transferase-conjugated SAWADEE domain (GST–SHH1, amino acids 125–258) construct was generated in the pENTR/

TEV/D plasmid (Invitrogen), recombined into the pDEST 15 plasmid (Invitrogen) and transformed into the Rosetta 2 (DE3) bacterial cell line (Novagen). Protein expression was induced by the addition of 500 µl of 1 M IPTG per 500 ml at $D_{600\,\mathrm{nm}}$ of 0.6 and cultures were grown at 16 °C overnight. At the time of induction the media was supplemented with 500 µl of 500 mM $ZnSO_4$. The GST fusion protein was then purified as described in ref. 34 and dialysed into storage buffer (50 mM Tris, pH 6.8, 300 mM NaCl, 40% glycerol, 2 mM DTT, 0.1% Triton X-100). The purified GST–SHH1 (125–258) protein was used to probe a MODified Histone Peptide Array (Active Motif) under the following conditions: The array was blocked at 25 °C for 45 min in a 5% milk 1× TBS solution, washed three times in a 1× TBS-T solution at 25 °C for 5 min, and then probed overnight at 4 °C with the GST–SHH1 SAWADEE domain protein at a concentration of 6.5 µg ml^{-1} in binding buffer (50 mM HEPES, pH 7.5, 50 mM NaCl, 5% glycerol, 0.4 mg ml^{-1} BSA, 2 mM DTT). The array was then washed three times as above, and probed an HRP conjugated GST antibody at a 1:5,000 dilution at 25 °C for 1 h. The array then washed as detailed above and developed using an ECL Plus kit (GE Healthcare).

Plant lines, site-directed mutagenesis, southern and western blotting. The various previously characterized *Arabidopsis* RdDM mutant alleles, the complementing SHH1–3×Myc–BLRP transgenic plant line, and the *pSHH1::SHH1–3×Myc–BLRP* construct used are as described in ref. 3. The *pol-iv* and *pol-v* mutants correspond to mutations in the *nrpd1* and *nrpe1* subunits of these polymerases, respectively. The structure-based mutations were generated in the *pSHH1::SHH1–3×Myc–BLRP* construct using a QuikChange Site Directed Mutagenesis kit (Stratagene) and were transformed into the *shh1-1* mutant background via the floral dip method. siRNA-seq and ChIP-seq experiments in the Col and RdDM mutant lines were conducted using floral tissue and BS-seq experiments were conducted using 10-day-old seedlings. Southern and western blotting experiments were conducted using tissue from the same individual plant lines in the T_1 generation and using previously described probes[35] and antibodies[36]. The siRNA-seq and BS-seq experiments in the SAWADEE domain point mutant lines were conducted using floral tissue or 10-day-old seedlings, respectively, from T_3 plants homozygous for the various *pSHH1::SHH1–3×Myc–BLRP* transgenes. The Pol-IV ChIP experiments and co-immunoprecipitation experiments in the various SAWADEE domain point mutant backgrounds were conducted using floral tissue from F_1 plants that were homozygous for the *shh1* mutant allele.

25. Feng, S., Rubbi, L., Jacobsen, S. E. & Pellegrini, M. Determining DNA methylation profiles using sequencing. *Methods Mol. Biol.* **733,** 223–238 (2011).
26. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
27. Chen, P. Y., Cokus, S. J. & Pellegrini, M. B. S. Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11,** 203 (2010).
28. Heisel, S. E. *et al.* Characterization of unique small RNA populations from rice grain. *PLoS ONE* **3,** e2871 (2008).
29. Spyrou, C., Stark, R., Lynch, A. G. & Tavare, S. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics* **10,** 299 (2009).
30. Cairns, J. *et al.* BayesPeak–an R package for analysing ChIP-seq data. *Bioinformatics* **27,** 713–714 (2011).
31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
32. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
33. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
34. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
35. Johnson, L. M., Law, J. A., Khattar, A., Henderson, I. R. & Jacobsen, S. E. SRA-domain proteins required for DRM2-mediated de novo DNA methylation. *PLoS Genet.* **4,** e1000280 (2008).
36. Law, J. A. *et al.* A protein complex required for polymerase V transcripts and RNA- directed DNA methylation in *Arabidopsis*. *Curr. Biol.* **20,** 951–956 (2010).

doi:10.1038/nature12294

Modulation of allostery by protein intrinsic disorder

Allan Chris M. Ferreon[1]*, Josephine C. Ferreon[1]*, Peter E. Wright[1,2] & Ashok A. Deniz[1]

Allostery is an intrinsic property of many globular proteins and enzymes that is indispensable for cellular regulatory and feedback mechanisms. Recent theoretical[1] and empirical[2] observations indicate that allostery is also manifest in intrinsically disordered proteins, which account for a substantial proportion of the proteome[3,4]. Many intrinsically disordered proteins are promiscuous binders that interact with multiple partners and frequently function as molecular hubs in protein interaction networks. The adenovirus early region 1A (E1A) oncoprotein is a prime example of a molecular hub intrinsically disordered protein[5]. E1A can induce marked epigenetic reprogramming of the cell within hours after infection, through interactions with a diverse set of partners that include key host regulators such as the general transcriptional coactivator CREB binding protein (CBP), its paralogue p300, and the retinoblastoma protein (pRb; also called RB1)[6,7]. Little is known about the allosteric effects at play in E1A–CBP–pRb interactions, or more generally in hub intrinsically disordered protein interaction networks. Here we used single-molecule fluorescence resonance energy transfer (smFRET) to study coupled binding and folding processes in the ternary E1A system. The low concentrations used in these high-sensitivity experiments proved to be essential for these studies, which are challenging owing to a combination of E1A aggregation propensity and high-affinity binding interactions. Our data revealed that E1A–CBP–pRb interactions have either positive or negative cooperativity, depending on the available E1A interaction sites. This striking cooperativity switch enables fine-tuning of the thermodynamic accessibility of the ternary versus binary E1A complexes, and may permit a context-specific tuning of associated downstream signalling outputs. Such a modulation of allosteric interactions is probably a common mechanism in molecular hub intrinsically disordered protein function.

Binding promiscuity is a hallmark of most hub proteins involved in signalling networks (for example, p53 and BRCA1)[8]. The inherent flexibility and structural adaptability of intrinsically disordered proteins makes them ideal hub proteins for binding to diverse partners. Not surprisingly, viruses widely use intrinsically disordered linear motifs to orchestrate subversion of the host cellular interactome[9].

The intrinsically disordered adenoviral protein E1A uses its amino-terminal region and conserved regions CR1 (residues 42–83) and CR2 (residues 121–139) in a cooperative manner to recruit numerous cellular regulatory proteins, thereby subverting signalling pathways in the infected cell[10]. The TAZ2 domain of CBP/p300 and the pocket domain of pRb each bind to two non-contiguous and largely non-overlapping regions of E1A to form binary complexes (E1A–pRb and E1A–TAZ2) and a ternary complex (pRb–E1A–TAZ2) (Fig. 1a)[11]. The major interaction site of CBP/p300 TAZ2 is within the E1A CR1 region, with a secondary binding site in the N-terminal region of E1A (Fig. 1b)[11]. pRb binds the characteristic LXCXE motif (residues 122–126) within the E1A CR2 region as well as a second binding site within CR1 (residues 42–49), in a region immediately preceding the TAZ2 binding site[12]. The E1A interaction sites occupy different regions of the CBP TAZ2 or the pRb surface[11,12]. The TAZ2 domain does not bind directly to the pocket domain of pRb, but rather associates with pRb only within ternary complexes formed by binding of both proteins to E1A[11]. To identify potential allosteric effects that fine-tune the interactions between the three proteins and assess the energetic contributions of each E1A interaction motif (N terminus, CR1 and CR2) to binding, several truncated E1A constructs were generated and studied (Fig. 1b, Methods).

Previous attempts at measuring dissociation constants (K_d) for E1A complexes with CBP by isothermal titration calorimetry (ITC) and nuclear

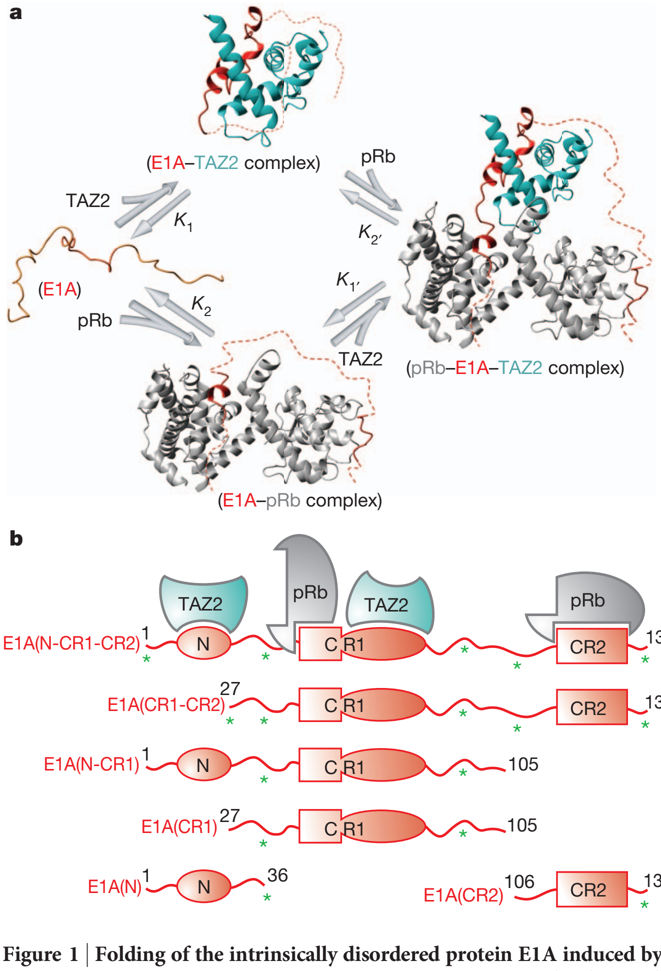


Figure 1 | Folding of the intrinsically disordered protein E1A induced by binding to pRb and the TAZ2 domain of CBP/p300. a, E1A binding and folding equilibria, showing formation of the ternary complex from the unbound intrinsically disordered protein state by way of two binary intermediate complexes. **b**, E1A constructs used to study the contributions of the N-terminal, CR1 and CR2 regions to formation of the binary and ternary complexes. Asterisks indicate the locations of single- or dual-site dye labelling for fluorescence measurements (that is, residue positions −3, 36, 88, 111 and 137, where residues 1–139 comprise the E1A sequence and positions −4 to −1 are the residues GSHM).

[1] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. [2] Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.
*These authors contributed equally to this work.

magnetic resonance (NMR) failed because E1A is highly aggregation-prone[11]. Even at concentrations as low as 10 μM, E1A(N-CR1-CR2) (residues 1–139 comprising the N-terminal region, CR1 and CR2) forms visible precipitates upon binding to CBP TAZ2. Hence, the previous NMR experiments, performed at micromolar concentrations, could demonstrate ternary complex formation between pRb, E1A and the CBP TAZ2 domain only for the short E1A CR1 region (E1A(CR1); residues 27–91)[11]. To overcome these problems, ensemble fluorescence anisotropy measurements were first attempted to measure binding affinities for longer E1A constructs that include more, or all, of the CBP, TAZ2 and pRb binding sites, and under more physiological concentrations (nanomolar to micromolar).

Formation of binary and ternary complexes of Ad2 E1A with the TAZ2 domain of mouse CBP and the human pRb pocket domain was monitored by fluorescence anisotropy titrations (Fig. 2a, b and Supplementary Fig. 1). From these ensemble fluorescence measurements we were able to obtain accurate dissociation constants for binary complexes with K_d greater than 25 nM (Supplementary Table 1 and Supplementary Fig. 2) that were in agreement with published values for pRb[12]. However, two issues impeded quantitative analysis of the binding data. First, most of the affinities are very high ($K_d < 25$ nM), and outside the reliable detection limit of fluorescence anisotropy measurements (Methods, Supplementary Table 1 and Supplementary Fig. 2). Second, in the presence of the N-terminal region, and especially for E1A(N-CR1-CR2) (residues 1–139), aggregation of the E1A constructs occurred at the relatively high concentrations required for competition fluorescence anisotropy assays.

To overcome the aforementioned problems, we used single-molecule fluorescence resonance energy transfer (smFRET). Owing to its high detection sensitivity, smFRET is an ideal method for investigating aggregation-prone and high-affinity systems, using low concentrations of fluorescently labelled protein (that is, ≤100 pM). In addition, the absence of ensemble averaging enables direct observation of free and bound populations, allowing for straightforward K_d measurements (Methods).

Binding affinities were measured by monitoring changes in intramolecular FRET that accompany folding of E1A upon binding to CBP TAZ2 or pRb. smFRET measurements were performed using freely diffusing E1A dual-labelled with donor and acceptor dyes (Fig. 1b). Free and bound populations in the resulting smFRET histograms (Fig. 3a–d and Supplementary Figs 3–6) have characteristic FRET efficiencies (E_{FRET}) (Supplementary Table 3) that are related to inter-dye distances[13,14]. In its free unbound state, E1A (labelled at multiple donor-acceptor sites) exhibited relatively low E_{FRET} (0.2–0.5), indicating extended structures. In contrast, E1A exhibited higher E_{FRET} (0.4–0.9) for both binary and ternary complexes, consistent with formation of more compact structures due to folding upon binding. The distances estimated from E_{FRET} values for bound E1A(CR1) are in agreement with those estimated from NMR and X-ray structures of E1A bound to TAZ2 and pRb pocket domains[11,12,15].

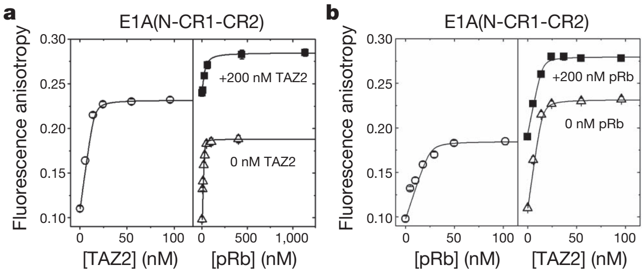


Figure 2 | E1A–TAZ2–pRb ternary complex formation detected by ensemble fluorescence anisotropy. **a**, **b**, TAZ2/pRb titration of free (open symbols) and TAZ2- or pRb-bound (solid symbols) Alexa Fluor 594-labelled E1A(N-CR1-CR2(1–139/S88C)).

Using E1A(CR1(27–105/36C88C)), with fluorescent labels attached at Cys introduced at residue positions 36 and 88 (Fig. 1b), titrations with increasing concentrations of CBP TAZ2 resulted in a gradual disappearance of the free E1A peak (E_{FRET} ~0.46), concurrent with a gradual appearance of the binary E1A–CBP TAZ2 peak (E_{FRET} ~0.9; Fig. 3a). The increased E_{FRET} of the latter peak is due to folding of E1A upon binding to TAZ2, forming a more compact E1A structure as observed by NMR[11]. The smFRET data can be fitted to a one-site binding model with a $K_d = 11.7 \pm 0.4$ nM (Methods and Supplementary Fig. 3). In the presence of ≥1 μM pRb, TAZ2 binds to the pRb-bound E1A(CR1(27–105/36C88C)) with $K_d = 37 \pm 6$ nM to form a ternary complex (Fig. 3b and Supplementary Fig. 3).

Next, we performed similar experiments using the more aggregation-prone E1A construct containing both the N-terminal region (residues 1–26) and the CR1 region (E1A(N-CR1(1–105/36C88C))). The observed E_{FRET} values for free and bound E1A(N-CR1) were very similar to those for E1A(CR1), suggesting that the CR1 region, located between the fluorescent probes, adopts similar conformations in both complexes, unperturbed by the presence of the N terminus. The N-terminal region of E1A seems to interact weakly with TAZ2, as the binding affinity increases ~4-fold when it is present (Supplementary Table 2). When the N-terminal region of E1A is free to participate in the binding interactions, $K_d = 3.2 \pm 0.5$ nM for TAZ2 binding to E1A alone, whereas $K_d = 1.5 \pm 0.3$ nM for binding of TAZ2 to E1A in the presence of 1 μM pRb (Fig. 3c, d and Supplementary Fig. 3). Binding of TAZ2 to the binary E1A(N-CR1)–pRb complex is much stronger than to the E1A(CR1)–pRb complex that lacks the E1A N terminus ($K_d = 1.5$ versus 37 nM), showing that the N-terminal region of E1A makes interactions that stabilize the ternary complex. Similar results were obtained for E1A constructs containing the CR2 motif (E1A(CR1-CR2(27–139/36C88C)) and E1A(N-CR1-CR2(1–139/36C88C))), where the E1A N terminus enhances the binding affinity for TAZ2 ($K_d = 1.6$ versus 7.5 nM for the shorter construct) but has no effect on binding to pRb (Supplementary Fig. 4 and Supplementary Table 2). This increase in affinity is attributed to additional binding interactions mediated by the E1A N terminus, which binds dynamically to a surface of TAZ2 opposite the CR1 binding site and causes exchange broadening of NMR resonances[11].

We next used the affinity data to generate protein phase diagrams (Fig. 3e–h), which provide graphical representations of folding and binding linkage equilibria[16]. These diagrams provide population information for different E1A species (free, binary and ternary species) versus concentrations of CBP TAZ2 and pRb. Each phase separation line (for example, black line, Fig. 3e–h) represents ligand concentrations where a corresponding state (for example, unbound E1A) is 50% populated relative to all other states. In cells, concentrations of signalling proteins range from nanomolar to micromolar and can be as high as millimolar with co-localization[17]. Therefore, the concentration ranges shown for the phase diagrams are well within physiological ranges. Asymmetry in the central white areas (where the population of none of the states exceeds 50%) reflects cooperative binding. Thus, for E1A(CR1), the decrease in TAZ2–E1A binding affinity in the presence of pRb and corresponding positive slope in the white area in Fig. 3e demonstrate negative cooperativity between pRb and CBP TAZ2, with the formation of the binary E1A complexes favoured over the ternary complex at lower concentrations of pRb and TAZ2. Notably, and in contrast with E1A(CR1), the E1A(N-CR1) binding phase diagram (Fig. 3f) reveals positive cooperativity for the interactions between CBP TAZ2 and pRb, clearly reflected in a negative slope of the white area. Therefore, the availability of the E1A N-terminal region can modulate the sign of the cooperativity of CBP TAZ2 and pRb binding to E1A(CR1). In the cell, this situation might have a key role when a binding partner sequesters the E1A CR2 region. Our observations are also directly relevant to the interactions of cellular proteins with CR2-deleted E1A produced by oncolytic adenovirus mutants that are in clinical trials for cancer therapy[18]. We note that truncated versions of E1A, lacking the N

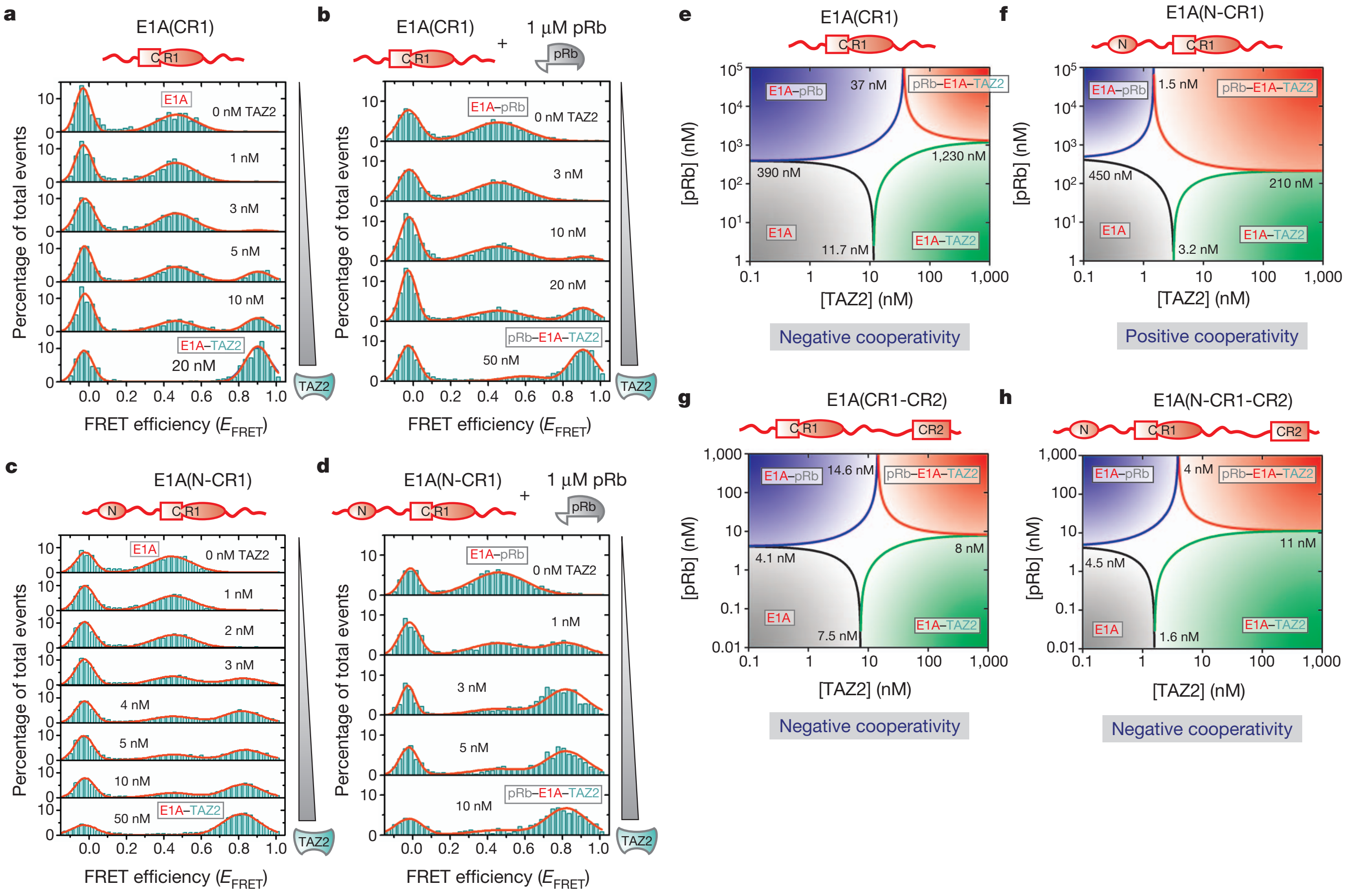


Figure 3 | E1A–TAZ2–pRb allosteric interactions probed using single-molecule fluorescence resonance energy transfer. a–d, smFRET histograms for the TAZ2 titration of Alexa Fluor 488- and 594-labelled E1A(CR1(27–105/36C88C)) (**a**, **b**) and E1A(N-CR1(1–105/36C88C)) (**c**, **d**) constructs in the absence (**a**, **c**) and presence (**b**, **d**) of 1 μM pRb. **e–h**, pRb–TAZ2 phase diagrams for FRET-labelled E1A(CR1) (**e**), E1A(N-CR1) (**f**), E1A(CR1-CR2(27–139/36C88C)) (**g**) and E1A(N-CR1-CR2(1–139/36C88C)) (**h**) constructed using the K_d values derived from ensemble and single-molecule fluorescence measurements (Supplementary Tables 1 and 2). The K_d values for the binding of E1A with pRb in the presence of CBP TAZ2 ($K_{2'}$) cannot be determined experimentally owing to overlap of E_{FRET} signals but can be calculated from a thermodynamic cycle analysis (Fig. 1a); $K_{1'}/K_1 = K_{2'}/K_2$. These values correspond to 1,230, 210, 8 and 11 nM in **e–h**, respectively.

terminus or other interaction domains, are commonly used to study the cellular response to E1A (ref. 5).

Previous NMR data suggest a plausible molecular basis for the observed negative cooperativity (Fig. 3e, g, h). Chemical shift titrations[11] indicate that binding of pRb disrupts a small subset of the intermolecular interactions that exist in the binary E1A(CR1)–CBP TAZ2 complex, indicating that negative allostery may be associated with partial overlap between the pRb and CBP TAZ2 binding sites in the E1A CR1 region. The molecular origin of the positive cooperativity observed for E1A (N-CR1) (Fig. 3f) is less obvious. However, allosteric coupling between sites in intrinsically disordered proteins can be either positive or negative and does not require mechanical linkage but can arise through perturbations of the energetic balance by binding events at individual sites[19].

The phase diagrams also provide a direct visualization of how cooperativity affects E1A population distributions and thereby their functional outcomes. Negative cooperativity (Fig. 3e, g, h) results in the ternary complex occupying a smaller area relative to the binary complexes. Conversely, positive cooperativity (Fig. 3f) results in the ternary complex occupying broader concentration ranges. Together, the phase diagrams demonstrate how multiple layers of regulation can be imposed on the E1A hub, depending on which domains of E1A are available for interaction with CBP/p300 and pRb, permitting the cooperativity of the system to be fine-tuned over a broad concentration range.

So far, there are relatively few examples of negatively cooperative biological systems[20]. Positive cooperativity is a common mechanism for increasing the binding potential. Positive cooperativity in ternary complex formation would enhance a critical function of E1A: the CBP/p300-mediated acetylation of pRb to force permanent exit from the cell cycle and promote differentiation of the host cell[21–23]. However, for a promiscuous molecular hub intrinsically disordered protein such as E1A (Fig. 4a), negative cooperativity has an equally important role because it broadens the stimulus range[24], increasing the population of intermediate binding states (binary complexes) and facilitating their interactions with other partners (Fig. 4b). This would permit a context-dependent modulation of different molecular species that contribute to the potency of viral E1A in subverting host cellular mechanisms[6,7].

Our results indicate that intrinsically disordered protein systems can be tuned to optimize population distributions and cellular outcome by changing the available binding sites. This could occur by competition between different molecular partners for the same E1A binding sites (Fig. 4a), resulting in allosteric modulation of the interaction and signalling networks involving CBP/p300 and pRb (Fig. 4b). E1A exhibits multiple activities in infected cells, mediating CBP/p300-dependent pathways that are independent of pRb (transcriptional activation or repression; green quadrant in Fig. 4b), pRb-dependent pathways that are independent of CBP/p300 (cell cycle progression, blue quadrant in Fig. 4b), and pathways that are dependent on both CBP/p300 and pRb (differentiation-specific functions of E1A, red quadrant in Fig. 4b)[10,21–23,25,26]. Our allosteric interaction modulation model provides an important mechanistic paradigm for understanding

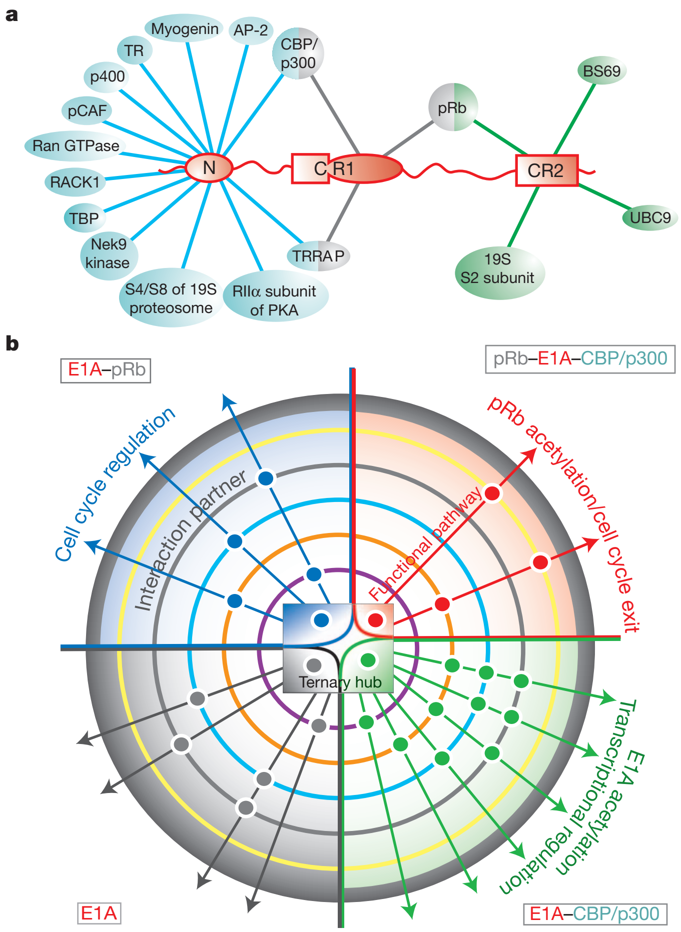


Figure 4 | E1A functional complexity achieved through binding promiscuity. a, Interactions of the N-terminal, CR1 and CR2 motifs of E1A with cellular proteins. Interactions mediated by the CR3 and CR4 regions of E1A[5] are not shown. **b**, Allosteric modulation of signalling pathways by interactions of the E1A–CBP/p300–pRb ternary hub. This hub, represented by a central phase diagram, has four E1A states: free E1A, E1A–pRb, E1A–CBP/p300, and ternary complex (grey, blue, green and red quadrants, respectively). Coloured concentric circles surrounding the hub represent additional protein partners with different interaction propensities for individual hub states. Each positive interaction is represented by a dot, coloured by hub state, and positioned based on the interaction partner. These ternary hub interactions with different sets of partners result in multiple functional pathways, the control of which may be achieved by modulating the central E1A–CBP/p300–pRb hub equilibria.

regulation of such varying signalling outputs, although other parameters are also likely to be important in a cellular context. A recent theoretical study[27] showed that allosteric ensembles associated with intrinsic protein disorder can upregulate or downregulate activity in response to different physiological stimuli, a feature of E1A that both activates and represses gene expression[10]. The capacity to expand protein functionality through modulation of accessible interaction sites has some similarities to alternative splicing, where the different protein isoforms generated increase the functional complexity of the genome[28]. Given the small size of the viral genome, it would be advantageous for adenovirus to amplify functional complexity using only a small number of proteins while maintaining the potential for maximum cellular control. A way to accomplish this is through a hub intrinsically disordered protein such as E1A that initially interacts with a small number of major binding partners (for example, pRb and CBP TAZ2) to form a series of hub interaction complexes (Fig. 4b). Additional binding partners then interact with the hub complex, with varied interaction preferences against different molecular forms, resulting in altered signalling outputs. Thus, modulation of allostery using intrinsically disordered protein regions that can bind to diverse partners may be a mechanism by which a promiscuous molecular hub intrinsically disordered protein can manage its functional complexity.

METHODS SUMMARY

K_d measurements using ensemble fluorescence anisotropy for the different E1A samples were performed using direct and competition titration methods[29].

The smFRET confocal instrumentation, and single-molecule data collection and analyses, were previously described[30]. For each smFRET measurement carried out at specific experimental conditions, an average of ~5,000 events were collected. For the whole series of smFRET titrations, data collection was in excess of 700,000 events. The E_{FRET} histograms were fitted to Gaussian functions using OriginPro 8 (OriginLab Corp.). K_d values were determined by using a one-site binding model and by analysing the ligand concentration dependence of the peak areas of free and bound populations. See Methods and Supplementary Fig. 7 for more details.

Full Methods and any associated references are available in the online version of the paper.

Received 16 November 2012; accepted 17 May 2013.

1. Hilser, V. J. & Thompson, E. B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl Acad. Sci. USA* **104,** 8311–8315 (2007).
2. Garcia-Pino, A. *et al.* Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity. *Cell* **142,** 101–111 (2010).
3. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6,** 197–208 (2005).
4. Xie, H. *et al.* Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6,** 1882–1898 (2007).
5. Pelka, P., Ablack, J. N., Fonseca, G. J., Yousef, A. F. & Mymryk, J. S. Intrinsic structural disorder in adenovirus E1A: a viral molecular hub linking multiple diverse processes. *J. Virol.* **82,** 7252–7263 (2008).
6. Ferrari, R. *et al.* Epigenetic reprogramming by adenovirus e1a. *Science* **321,** 1086–1088 (2008).
7. Horwitz, G. A. *et al.* Adenovirus small e1a alters global patterns of histone modification. *Science* **321,** 1084–1085 (2008).
8. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D^2 concept. *Annu. Rev. Biophys.* **37,** 215–246 (2008).
9. Davey, N. E., Trave, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **36,** 159–169 (2011).
10. Berk, A. J. Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus. *Oncogene* **24,** 7673–7685 (2005).
11. Ferreon, J. C., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Structural basis for subversion of cellular control mechanisms by the adenoviral E1A oncoprotein. *Proc. Natl Acad. Sci. USA* **106,** 13260–13265 (2009).
12. Liu, X. & Marmorstein, R. Structure of the retinoblastoma protein bound to adenovirus E1A reveals the molecular basis for viral oncoprotein inactivation of a tumor suppressor. *Genes Dev.* **21,** 2711–2716 (2007).
13. Ferreon, A. C. M., Moran, C. R., Gambin, Y. & Deniz, A. A. Single-molecule fluorescence studies of intrinsically disordered proteins. *Methods Enzymol.* **472,** 179–204 (2010).
14. Deniz, A. A., Mukhopadhyay, S. & Lemke, E. A. Single-molecule biophysics: at the interface of biology, physics and chemistry. *J. R. Soc. Interface* **5,** 15–45 (2008).
15. Lee, J.-O., Russo, A. A. & Pavletich, N. P. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature* **391,** 859–865 (1998).
16. Ferreon, A. C. M., Ferreon, J. C., Bolen, D. W. & Rösgen, J. Protein phase diagrams II: nonideal behavior of biochemical reactions in the presence of osmolytes. *Biophys. J.* **92,** 245–256 (2007).
17. Kuriyan, J. & Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **450,** 983–990 (2007).
18. Heise, C. *et al.* An adenovirus E1A mutant that demonstrates potent and selective systemic anti-tumoral efficacy. *Nature Med.* **6,** 1134–1139 (2000).
19. Hilser, V. J. & Thompson, E. B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl Acad. Sci. USA* **104,** 8311–8315 (2007).
20. Cui, Q. & Karplus, M. Allostery and cooperativity revisited. *Protein Sci.* **17,** 1295–1307 (2008).
21. Wang, H.-G. H., Moran, E. & Yaciuk, P. E1A promotes association between p300 and pRB in multimeric complexes required for normal biological activity. *J. Virol.* **69,** 7917–7924 (1995).
22. Chan, H. M., Krstic-Demonacos, M., Smith, L., Demonacos, C. & La Thangue, N. B. Acetylation control of the retinoblastoma tumour-suppressor protein. *Nature Cell Biol.* **3,** 667–674 (2001).

23. Nguyen, D. X., Baglia, L. A., Huang, S.-M., Baker, C. M. & McCance, D. J. Acetylation regulates the differentiation-specific functions of the retinoblastoma protein. *EMBO J.* **23,** 1609–1618 (2004).
24. Koshland, D. E. Jr. The structural basis of negative cooperativity: receptors and enzymes. *Curr. Opin. Struct. Biol.* **6,** 757–761 (1996).
25. Sang, N., Avantaggiati, M. L. & Giordano, A. Roles of p300, pocket proteins, and hTBP in E1A-mediated transcriptional regulation and inhibition of p53 transactivation activity. *J. Cell. Biochem.* **66,** 277–285 (1997).
26. Green, M., Panesar, N. K. & Loewenstein, P. M. The transcription-repression domain of the adenovirus E1A oncoprotein targets p300 at the promoter. *Oncogene* **27,** 4446–4455 (2008).
27. Motlagh, H. N. & Hilser, V. J. Agonism/antagonism switching in allosteric ensembles. *Proc. Natl Acad. Sci. USA* **109,** 4134–4139 (2012).
28. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genet.* **30,** 13–19 (2002).
29. Lee, C. W., Ferreon, J. C., Ferreon, A. C. M., Arai, M. A. & Wright, P. E. Graded enhancement of p53 binding to CBP/p300 by multisite phosphorylation. *Proc. Natl Acad. Sci. USA* **107,** 19290–19295 (2010).
30. Ferreon, A. C. M., Gambin, Y., Lemke, E. A. & Deniz, A. A. Interplay of α-synuclein binding and conformational switching probed by single-molecule fluorescence. *Proc. Natl Acad. Sci. USA* **106,** 5645–5650 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Manlapaz for technical support, A. Jansma and G. Bhabha for preparation of plasmid constructs, P. Haberz for mass spectrometry, and J. Dyson and M. Martinez-Yamout for discussions. This work was supported by grants GM066833 (A.A.D.) and CA96865 (P.E.W.) from the National Institutes of Health, and the Skaggs Institute for Chemical Biology.

Author Contributions A.C.M.F. and J.C.F. performed the experiments. A.C.M.F., J.C.F., P.E.W. and A.A.D. designed experiments, analysed data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A.D. (deniz@scripps.edu) or P.E.W. (wright@scripps.edu).

METHODS

Sample preparation. Protein expression and purification were carried out as described[11]. The Ad2 E1A short constructs (E1A(CR1(27–105)), E1A(N-CR1(1–105)) and E1A(CR2(106–139))) were obtained via thrombin digestion of the longer E1A constructs (E1A(CR1-CR2(27–139)) or E1A(N-CR1-CR2(1–139)))[11]. All E1A Cys mutants used for ensemble fluorescence[29] or single-molecule fluorescence resonance energy transfer (smFRET)[30] experiments have the additional mutations C6S and C124S, which replace two natural Cys residues that are respectively located in CBP TAZ2 and pRb binding regions. Although C124 is in a conserved pRb LXCXE binding motif, it has been shown that a Cys to Ser mutation in the site exhibits marginal effects on E1A binding[31]. Alexa Fluor 488 and 594 (Molecular Probes) fluorescent dyes were attached at sites that are unlikely to cause structural perturbations or affect E1A binding to CBP TAZ2 or pRb (residue positions −3, 36, 88, 111 and 137, where residues 1–139 comprise the E1A sequence and positions −4 to −1 are the residues GSHM).

For direct E1A titrations against CBP TAZ2 and/or pRb monitored by ensemble fluorescence anisotropy, E1A constructs with single Cys (S36C for E1A(N(1–36)); E137C for E1A(CR2(106–139)); otherwise, S88C) were used to attach Alexa Fluor 594 probes. For ensemble competition experiments, the competing E1A ligands (E1A(N-CR1-CR2(1–139)), E1A(CR1-CR2(27–139)), E1A(N(1–36)), and E1A(CR2(106–139))) have the wild-type E1A sequence except for the G139 residue that was mutated to Trp for more accurate protein concentration determination by ultraviolet spectroscopy. For E1A(CR2(106–139)), protein with the wild-type sequence was used in direct titration measurements, with the E1A protein N-terminally labelled with Dylight594 (Thermo Scientific) NHS ester probe. To investigate the role of the E1A N terminus in the protein's binding properties, four sets of pair-labelled E1A constructs were used for smFRET studies: E1A(CR1(27–105)) and E1A(N-CR1(1–105)), 36C88C; and, E1A(CR1-CR2(27–139)) and E1A(N-CR1-CR2(1–139)), 36C88C, (−3)C111C, 36C137C.

All E1A single Cys mutants were labelled in 50 mM Tris, 6 M guanidine HCl, pH 7.2 using ~3–5 fold molar excess of maleimide dye. For E1A double Cys mutants, approximately 5 nmol of E1A were incubated with 1:3 concentration ratio of Alexa Fluor 488: Alexa Fluor 594 dye. Labelling reactions were run for 2 h at room temperature. All dye-labelled E1A samples were purified using an analytical C18 reverse-phase HPLC column, and were checked for correct mass and for incorporation of the Alexa dyes by MALDI–TOF mass spectrometry.

Ensemble fluorescence spectroscopy. Isothermal titrations in 20 mM Tris, 50 mM NaCl, 1 mM dithiothreitol (DTT), pH 7.0 at 21 °C were performed by monitoring ensemble fluorescence anisotropy. Two titration methods were used: direct protein–ligand titration and competition binding measurements[29]. Direct titrations were carried out by detecting fluorescence anisotropy changes in solutions containing 25 nM of dye-labelled E1A macromolecule (M) as a function of ligand (L) concentration (CBP TAZ2 or pRb). Dissociation constants (K_d) were determined using OriginPro 8.0 (OriginLab Corp.) by nonlinear least-squares (NLS) fitting of the data to a one-site binding model (see equation (2) below). To determine the goodness of fit and test the validity of the simplified model, simulations were performed using the fitted parameters and compared to the data on the basis of a more complete binding model that considers the macromolecule concentration (see the model defined below and described by equation (3)). Application of the more exact model is not feasible for the analysis of the ensemble fluorescence anisotropy data due to the number of fitting parameters. For cases of low-affinity binding, where $K_d \gg M$ (such as with the titration of E1A(CR1(27–105)) and E1A(CR2(106–139)) against pRb), the assumptions of the simplified one-site binding model become valid, as can be shown by simulations using the derived parameters as applied to the second model. For cases of high-affinity binding, where $K_d \ll M$, the estimates for K_d using the first binding model are not accurate. In such cases, an upper bound for the K_d was used (Supplementary Table 1). For the competition method, 25 nM labelled E1A were initially bound with 500–1,000 nM pRb or 350–500 nM CBP TAZ2, and competed with the unlabelled E1A counterpart to see the effect of the probes (for example, E1A(N-CR1(1–105)), 88C-Alexa Fluor594, competed against wild-type sequence E1A(N-CR1(1–105))). An estimate of the K_d from the direct titration was necessary to fit the K_d of the competing ligand.

Single-molecule spectroscopy. Single-molecule smFRET experiments were carried out as described previously[30] using a home-built laser confocal microscope system that uses an Axiovert 200 microscope (Zeiss). Excitation was achieved by focusing the 488-nm line of a 543-AP-A01 tunable argon-ion laser (Melles Griot) into the sample solution, 30 µm above a glass coverslip surface, using a water immersion objective (1.2 NA, ×63; Zeiss). The fluorescence emission was collected using the same objective, separated from the excitation light using a dichroic mirror (Q495LP; Chroma Tech. Corp.), spatially filtered using a 100 µm-pinhole then split into donor and acceptor components using a second dichroic mirror (560 DCXR; Chroma). The donor and acceptor signals were further filtered using an HQ 525/50M band-pass filter (donor; Chroma) and a 590 LPV2 long-pass filter (acceptor; Chroma), then detected using SPCM-AQR-14 avalanche photodiode (APD) photon counting modules (Perkin-Elmer Optoelectronics). Photon counts were recorded using a photon counting card (PCI 6602; National Instruments) interfaced with a computer.

FRET efficiency (E_{FRET}) histograms were generated by using a two-channel data collection mode to simultaneously record donor and acceptor signals as a function of time, with a binning time of 500 µs. The donor–acceptor solutions used were ~100 pM in fluorophore concentration (that is, ~100 pM FRET-labelled E1A), ensuring that virtually all of the detected signals were from single molecules. The background counts, the leakage of donor emission into the acceptor channel (~8%) and the acceptor emission due to direct excitation (~5%) were determined in separate experiments, and used to correct the signals before FRET analysis. A threshold of 50 counts (the sum of signals from the two channels) was used to separate background noise from fluorescence signals, and E_{FRET} values were calculated for each accepted event using equation (1) and plotted in the form of histograms:

$$E_{\mathrm{FRET}} = I_A/(I_A + \gamma I_D) \qquad (1)$$

I_D and I_A are the corrected donor and acceptor fluorescence intensities, respectively, and γ is a correction factor that is dependent on the donor and acceptor fluorescence quantum yields, and donor channel and acceptor channel detection efficiencies. Using the same experimental setup and FRET dye-pair, we previously measured γ to be approximately equal to unity[30]. Although the accuracy of the determined γ value is critical for measurement of inter-dye distances, γ does not have a part in the calculation of population distributions[30].

Direct detection of binding events using smFRET. Binding of unlabelled CBP TAZ2 and/or pRb to different constructs of E1A labelled with Alexa Fluor 488 (donor) and 594 (acceptor) (see 'Sample preparation' section above and Fig. 1b) was detected using smFRET at room temperature (~21 °C). The same solution conditions were used as for the ensemble fluorescence measurements (20 mM Tris, 50 mM NaCl, 1 mM DTT, pH 7.0). An average of ~5,000 single-molecule events was collected for each smFRET histogram measurement. In total, the complete set of smFRET titration data reported here comprise in excess of 700,000 events. Representative smFRET histograms are shown in Fig. 3a–d and Supplementary Figs 3–6.

E_{FRET} histograms were fitted to Gaussian functions by using OriginPro 8.0 with the peak positions, areas and widths at half height used as fitting parameters. For experimental conditions where E1A predominantly adopts a single binding state (that is, free, CBP TAZ2-bound, pRb-bound, or in ternary complex with CBP TAZ2 and pRb), smFRET histograms showed two peaks: one corresponding to the protein signal, and another to the 'zero peak', which is present in all histograms due to molecules with photo-bleached, missing, or non-fluorescent acceptor probe. These histograms of single populations or 'pure states' were used as references in determining (via independent NLS Gaussian fits) the characteristic E_{FRET} signatures of the different E1A binding states (see Supplementary Table 3). These precisely determined E_{FRET} values were then used as fixed parameter inputs in the analyses of smFRET histograms exhibiting resolved multiple 'protein peaks' that correspond to different E1A conformations (for example, unbound and CBP TAZ2-bound states). The areas under each protein peak determined by NLS Gaussian fitting were then used to calculate fractional populations (for example, fraction unbound and fraction CBP TAZ2-bound), which were analysed further as a function of ligand concentration (for example, [CBP TAZ2]) to determine binding constants (see Fig. 1a and Supplementary Table 2, and K_d determination method discussion below). The smFRET data presented in Supplementary Figs 3a–f, 4a–f, 5a–c and 6a–d were all analysed independently via NLS Gaussian fitting.

In some cases, smFRET histograms acquired under different solution conditions were analysed simultaneously, sharing fitting parameters that correspond to the same protein states. This global analysis was especially useful in cases where protein peaks were not resolved well or when E_{FRET} values cannot be satisfactorily determined independently using just the histograms of pure states. Fractional populations were calculated using the area parameters derived from global fitting and analysed further for K_d determination (see below). The smFRET data presented in Supplementary Fig. 5d–f were analysed via global NLS Gaussian fitting.

K_d determination by smFRET. Detection of macromolecular interactions in solution at single-molecule resolution holds a number of important advantages over ensemble methods, including the direct measurement of population distributions, the ability to experimentally handle aggregation-prone systems, and improved resolution in the study of high-affinity interactions.

Applying smFRET to derive K_d values for the binding of a ligand L to a macromolecule M, assuming that the bound and unbound populations of the FRET-labelled macromolecule exhibit distinct FRET efficiencies, is straightforward and can be performed empirically, without the need for model fitting. K_d, or the ligand concentration at which the bound (ML) and unbound (M) macromolecule populations are equal, can be determined simply by titrating M with increasing concentration of L until the measured smFRET histogram shows approximately equal areas for the peaks associated with the two binding states. The process can then be repeated several times to achieve the desired precision.

Alternatively, population distributions can be similarly measured, then used as an experimental variable that depends on $[L]$ and analysed using a binding model. The model described by equation (2) (see below) assumes that the total ligand concentration $[L_T]$ is approximately equal to the concentration of free ligand, that is, $[M] \ll K_d$, a requirement that is easily achieved using smFRET, in which measurements are usually performed using 100 pM labelled molecules (or less).

Binding constants for the E1A–CBP TAZ2–pRb ternary system (Fig. 1a) were determined as a function of ligand concentration (that is, [CBP TAZ2] or [pRb]) using the fractional populations (that is, fraction bound and unbound) directly derived from the analyses of smFRET histogram data (described above). Fraction populations plotted against total ligand concentration (expressed in terms of $-\log[\text{ligand}]_{\text{total}}$ or $\text{p}L_T$) were analysed graphically using the general binding model, $ML \leftrightarrow M + L$, and fitted with OriginPro 8.0 using equation (2):

$$Y = (\alpha Y_M + Y_{ML})/(1 + \alpha) \qquad (2)$$

M represents a macromolecule binding to a ligand L, Y is the experimental observable (that is, fraction bound or unbound), Y_M and Y_{ML} are the binding transition baselines (that is, the constants 0 and 1, respectively, if using fraction bound as Y; otherwise, 1 and 0), $\alpha = 10^{\zeta}$, $\zeta = \text{p}L_T - \text{p}K_d$, $[\text{ligand}]_{\text{total}}$ represents both bound and unbound forms of L, $\text{p}K_d = -\log[K_d]$, and K_d is the dissociation constant. The model assumes that the concentration of unbound ligand is approximately equal to $[\text{ligand}]_{\text{total}}$, that is, the total concentration of the macromolecule E1A (~100 pM) is significantly less than the K_d values being measured, which in the case here are in the 1–50 nM range (see Supplementary Table 2). In addition, Y can be any observable/signal that is able to distinguish the different binding states, for example, E_{FRET} (see Supplementary Fig. 5c).

A more general expression describing the same model (that is, $ML \leftrightarrow M + L$) is given by equation (3):

$$Y = ([ML]/[M_T])(Y_{ML} - Y_M) + Y_M \qquad (3)$$

M_T is the total M concentration independent of ligation state, Y is the observable, Y_M and Y_{ML} are the binding transition baselines, and $[ML] = (-b - (b^2 - 4ac)^{0.5})/2a$, with $a = 1$, $b = -K_d - [M_T] - [L_T]$, and $c = [M_T][L_T]$. Presented in Supplementary Fig. 7 are simulations for ligand binding at different M_T, highlighting the advantage of single-molecule detection in resolving binding constants of high-affinity interactions.

Protein phase diagrams. Using the binding constants derived from ensemble and single-molecule measurements (see above and Supplementary Tables 1 and 2), phase diagrams were generated to visualize the ligand concentration dependence of E1A interaction with its binding partners CBP TAZ2 and pRb (Figs 3e–h). Detailed descriptions of the general properties of protein phase diagrams, and their construction and interpretation, are provided elsewhere[16,32].

Here, we use the reaction mechanism depicted in Fig. 1a to describe the coupled folding and binding of E1A with CBP TAZ2 and pRb. K_1 and $K_{1'}$, and K_2 and $K_{2'}$ are equilibrium constants for E1A binding to CBP TAZ2 in the absence and presence of pRb, and to pRb in the absence and presence of CBP TAZ2. Because the reaction scheme constitutes a complete thermodynamic cycle, it can be shown that $K_1/K_{1'} = K_2/K_{2'}$. 50% phase separation lines were constructed as previously described[32], using partition functions (Q_i) that describe each of the four binding states (that is, unfolded and unbound (U), folded and CBP TAZ2-bound (FL_1), folded and pRb-bound (FL_2) and ternary (FL_1L_2) states). For example, 50% phase separation lines between the U state and the three other binding states are calculated by equating Q_U with the sum of the remaining partition functions Q_{FL1}, Q_{FL2} and Q_{FL1L2}.

31. Kadri, Z. *et al.* Direct binding of pRb/E2F-2 to GATA-1 regulates maturation and terminal cell division during erythropoiesis. *PLoS Biol.* **7,** 1–15 (2009).
32. Rösgen, J. & Hinz, H. J. Phase diagrams: a graphical representation of linkage relations. *J. Mol. Biol.* **328,** 255–271 (2003).

COLUMN Outreach works best when people stumble over research unawares **p.397**

CAREERS BLOG News and discussions on research jobs **go.nature.com/z8g4a7**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

RYCCIO/GETTY

MOBILE APPS

A conference in your pocket

Meeting attendees can use apps to network and ease logistical hassles.

BY ROBERTA KWOK

A tweet caught Jessica Ball's attention at last year's meeting of the American Geophysical Union (AGU). A panel had been added about the trial in which six Italian scientists had been found guilty of manslaughter for their handling of earthquake-risk communication, shortly before a magnitude-6.3 quake struck the city of L'Aquila in 2009, killing 309 people. Ball, a PhD student in geology at the University at Buffalo in New York, added the session to her itinerary using the iPad app for the AGU meeting in San Francisco, California. During the panel, she tweeted about speakers' key points, including the importance of separating the roles of science advisers and civil authorities.

Ball also met geologists and a science-communication expert for drinks during the conference — organized through Twitter. She used her tablet to show other attendees videos about the lava domes that she studies, and displayed a poster with Quick Response (QR) barcodes that allowed people to access the same videos online using their smartphones.

Conferences have come a long way in recent years. Attendees used to base their planning on phonebook-sized paper programmes that they lugged around in tote bags, and communicate only with people they happened to bump into at coffee breaks. Now, a host of apps on smartphones and tablets allows attendees to expand their networking, search meeting programmes, get schedule updates, discover under-the-radar events, share information and offer better explanations of their work. As long as attendees make sure that they don't spend the entire meeting glued to screens, mobile tools can facilitate lively online conversations, inform research and pave the way for face-to-face meetings.

FOLLOWING THE BUZZ

Twitter is by far the most popular channel for online conference chatter. The event's official hashtag can lead users to organizers, panellists and attendees already tweeting about the meeting. Tweets about an upcoming session might suggest whether it is worth ▶

▶ attending, and comments about an ongoing or completed panel allow people to pick up the main points if they couldn't attend. "It takes the stress out of feeling like you have to be everywhere at once," says Kelle Cruz, an astronomer at Hunter College in New York. Scientists also can track the buzz about their own talks by creating a hashtag specifically for their session.

Twitter is also a crucial networking tool, helping people to connect with fellow attendees who have similar interests. Users can invite Twitter connections for coffee or look out for their name tags at the conference, paving the way for an in-person introduction, says Emily Jane McTavish, an evolutionary biologist at the University of Kansas in Lawrence. "That's made a big difference to me at meetings where I didn't know people," she says. Jeremy Yoder, an evolutionary geneticist at the University of Minnesota in St Paul, used Twitter to help to organize a lunch for lesbian, gay, bisexual and transgender scientists at the First Joint Congress on Evolutionary Biology in Ottawa last year. And although these connections might not lead to immediate work advantages, one never knows who might be on one's next grant-review panel or job-search committee, says Cruz.

JACQUELINE FAHERTY

"It takes the stress out of feeling like you have to be everywhere at once."
Kelle Cruz

SAVE IT FOR LATER

People sometimes tweet details about sessions they attend, as a way of taking notes. Holly Bik, a marine genomicist at the University of California, Davis, finds that her notes are often too long-winded if she types them out in a word-processing programme, but Twitter's 140-character limit helps her to distil out the main points. She also can quickly add links to papers. Later, she uses Storify (a website that is also available as an iPad app) to collect and archive relevant tweets so that she can easily access them later.

Tweeting helps scientists who can't attend the conference to follow important developments, which is particularly appreciated at small meetings. Some people tweet to ask for clarification from other attendees during a talk — for example, to request help understanding a figure, or to find out what an acronym stands for. Users should make sure, however, that the conference doesn't have a policy against tweeting — and should be careful not to disrupt the presentation by talking on their phone, leaving the ringer on or typing ceaselessly. They should also take care to avoid tweeting or posting excessively harsh critiques of data or presentations, given that they can be seen by just about anybody online — including the speaker. Critiques in general are acceptable, but users need to be as diplomatic digitally as they would be in person.

Twitter can also be a good way to communicate with meeting organizers, who may be able to answer logistical questions such as where to eat or how to deal with problems with the audiovisual equipment. It can be less disruptive than ducking out of a session to make a phone call, and multiple organizers and attendees will be able to see the question, increasing the chances of a quick response. At the 2013 meeting of the American Association for the Advancement of Science (AAAS) in Boston, Massachusetts, in February, organizers responded to tweets sent to the official @AAASMeetings account within minutes, says Tiffany Lohwater, director of meetings and public engagement for the association in Washington DC.

Tweeting at conferences can even lead to unexpected career developments. McTavish once tweeted about the lack of other female attendees at a computer-science workshop. One of her Twitter followers happened to be organizing a computational phylogenetics hackathon — a meeting of biologists and programmers to develop new software tools — and invited McTavish to apply to attend. Her participation in the hackathon led to a paper and an opportunity to reconnect with the scientist who ultimately became her postdoc supervisor.

Other social-media tools may also help attendants to navigate the myriad sessions and plenary events at a conference. Organizations sometimes post meeting updates or highlights on their Facebook pages; users can 'Like' the page to see the updates in their news feed. And when the meeting is over, attendees can maintain connections by sending 'friend' requests through Facebook or the more professional LinkedIn.

CONFERENCE LOGISTICS

Increasingly, conference attendees can turn to apps that are tailored to specific meetings. The quality and features vary, but such apps often include schedules, abstracts, presenter biographies, PDF and PowerPoint files uploaded by speakers, venue maps, lists of nearby restaurants, and ways to take notes and save contacts. Some of them also work offline — a boon when the conference Wi-Fi gets bogged down. "You don't want to be sitting there waiting for pages to load," says Silke Fleischer, co-founder of ATIV Software in Santa Rosa, California, which develops event-planning and conference apps. Among others, it has provided app software for the 2012 Society for Neuroscience meeting in New Orleans, Louisiana, which had about 28,000 attendees.

Conference organizers can build their apps with various providers to get a range of features. ATIV's EventPilot app lets users view downloaded PowerPoint slides, which is useful if the projector quality is poor. Attendees can also exchange contact information by scanning QR barcodes on each other's phones, even without an Internet connection.

EventMobi in Toronto, Canada, provides a live polling feature that allows speakers to ask questions of the audience and get a real-time chart of the results. An app by Bizzabo in New York suggests attendees with similar interests for users to contact, and Bloodhound in San Francisco will soon offer a feature to look at sessions that users have chosen to attend, and suggests others that they might like.

Users can turn to other apps if the conference software doesn't offer the required features (see 'Appy to help'). And some apps can help researchers in the lab, too (see *Nature* **484,** 553–555; 2012).

Online tools and mobile devices have even infiltrated the old-fashioned poster session.

ON THE GO

Appy to help

Here are some of the most useful apps for conference-goers.

- Twitter, HootSuite and Echofon: read Twitter feeds, send tweets, follow other users and search for hashtags.
- Notability and iAnnotate PDF: type or handwrite notes on a PDF file, such as a conference programme. Some apps sync notes to cloud services such as Dropbox and Google Drive.
- Notes, Simplenote, and Evernote: take notes that are synchronized across devices.
- GoodNotes and Papyrus: handwrite notes or draw with a stylus.
- WorldCard Mobile and CamCard: scan business cards and automatically import the information into a phone's contacts list. Supports multiple languages.
- Scan: use Quick Response barcodes to view associated websites automatically. Keeps history of past scans.
- Yelp, Urbanspoon and OpenTable: search for nearby restaurants, read reviews and make reservations.
- Expensify and Concur: scan receipts and create expense reports. R.K.

Researchers can use websites such as Kaywa to generate QR barcodes to embed in their posters. Viewers can scan the barcodes with a smartphone or tablet to automatically open a web page showing videos, linked papers or further data.

Apps come in handy for hallway conversations, too. Ball uses Skitch and Paper to draw pictures with her finger or a stylus, illustrating concepts in her volcano research — the locations of hot springs, for example, or the direction of fluid movement in a lava dome. She can e-mail the pictures to others or save them as ideas for figures.

Carol Finn, president of the AGU, uses the Keynote app on her iPad to show slides from her presentation, and EarthObserver to demonstrate features of the area she is studying, such as topography. For Android users, Quickoffice Pro and Google Earth, respectively, perform some of the same functions.

In future, other conference interactions may also move into apps. The AGU is considering adding scoring forms for its student-paper competition — in which volunteer attendees judge students' presentations at poster sessions — to its app. It is also thinking about adding discussion boards on which people can ask presenters questions about uploaded posters or recorded talks. This year, Bizzabo will start offering polling so that registered users can vote on their favourite sessions.

Apps with indoor mapping might one day pinpoint attendees' location in the building and direct them to the next session on their itinerary. Organizers could make conferences into a game by giving people rewards for going to specific activities or booths.

Some meetings might soon drop paper programmes all together. The Association for the Sciences of Limnology and Oceanography in Waco, Texas, offered an app at its meeting for the first time in 2013, and will probably go paperless in a few years, says co-organizer Hans-Peter Grossart, a microbial ecologist at the Leibniz Institute of Freshwater Ecology and Inland Fisheries in Neuglobsow, Germany. However, the AGU and the AAAS plan to offer paper programmes for the foreseeable future.

Conference attendees do need to exercise caution when turning to the blizzard of digital tools. Taking photos of slides, or recording talks without the speaker's permission, is generally considered bad form. And users should try not to get distracted by the constant stream of tweets and notifications during real-life conversations. "You want to be present," says McTavish. Emma Borochoff, a community manager at Bizzabo, agrees. "Connections aren't complete if they're just online," she says. ■

Roberta Kwok *is a freelance science writer in Seattle, Washington.*

COLUMN

Roadside science

Sometimes the best outreach happens when lay people stumble over research unawares, says **Carolyn Beans**.

CLAIRE WELSH/NATURE

As government funding of science declines and public scepticism runs rampant, scientists are working hard to find effective ways to explain their research to others. Although social media, magazines, museums and nature centres are all important vehicles for bringing science to the public, I worry that they draw only people who already appreciate research. How can we reach those who are resistant or indifferent?

With this concern in mind, I take my job as a roadside scientist very seriously.

I study an invasive plant in the northeastern United States, and find myself working along roadways that encourage its spread. I wear red rain boots, a reflective vest, a waist pouch that I have fashioned into a tool belt, and a baseball cap draped with mosquito netting. I look … well, odd. Cyclists, pedestrians and drivers often stop to ask what on Earth I am doing.

In Lubec, Maine, an electrician parked and came over to check out my work. I pointed to the seedlings of the native and invasive jewelweeds that I study, then showed him the seedling that I was folding into my plant press. The leaves of my sample were coloured like those of the native jewelweed, but shaped more like those of the invasive species, suggesting that the sample was a hybrid. It made me wonder whether the invasive jewelweed might invade not only the native's space, but also its genome. The electrician marvelled that this was what science looked like — a woman on the side of the road folding plants in newsprint.

In Camden, Maine, a neighbour asked about the mesh bags that I was placing over developing fruits of the native jewelweed. I explained that I needed seeds from the native species living in places both with and without the invasive plant. I could grow these seeds in competition with the invasive species in a greenhouse, to test whether the native seedlings from invaded communities survived better than those from communities without the invaders. The neighbour was excited to learn that evolution could take place on his own street, and that it was actually measurable.

Sometimes I am tempted to brush off a curious passerby. As the field season wanes, any distraction feels as if it could result in enough data loss to ruin an experiment. But I have gained meaningful insights from my interactions with the public. An older woman in Camden who showed me the local children's library told me that she recalled seeing the magenta flowers of the invasive jewelweed in her neighbourhood as many as 50 years earlier. This suggested to me that the native jewelweed has had at least half a century to evolve in response to the invasion. In return, I told the woman that the plant she admired from her bedroom window was originally from India.

Not everyone who stops to ask what I am doing hangs around long enough to hear my response. Some are turned off by the mention of science. Others are too busy to chat. But if I explain my research clearly enough, many passersby want to hear more.

As a roadside scientist, I have the opportunity to talk to members of the public without first drawing them to a blog or museum. People who would never seek out a scientific discussion come to me unaware that we are about to talk about invasive species, evolution and what it is like to be a field biologist. I receive them not knowing whether they accept evolution, or if science is going to be a tough sell. I offer them an explanation of my work and they offer me the chance to win them over.

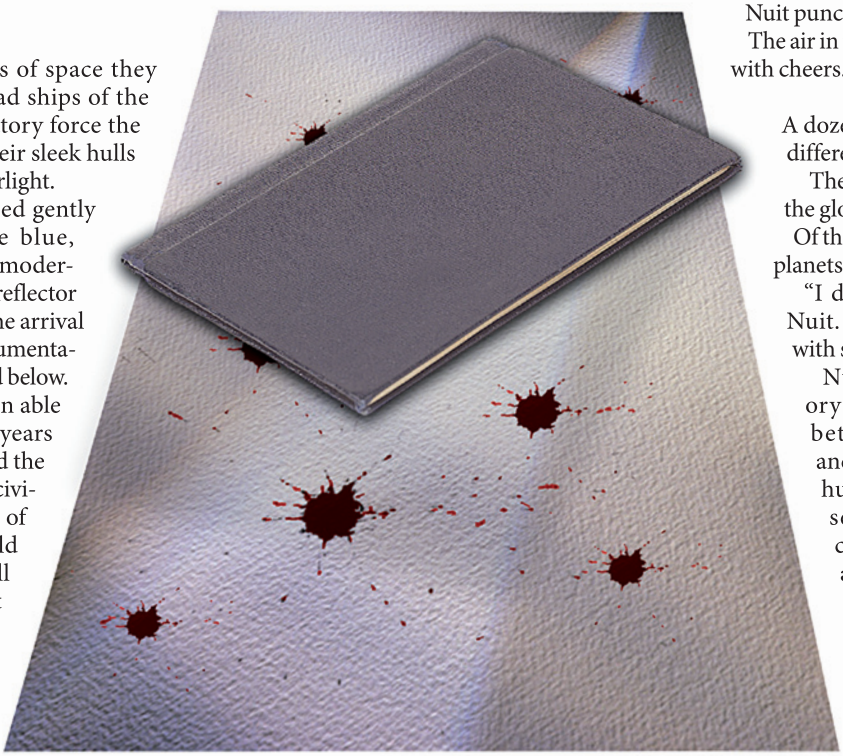
Just as successful political campaigns recognize that knocking on doors brings people to the polls, I believe that impromptu face-to-face communication brings people to an appreciation of science. In a way, we all become roadside scientists every time we describe our research to a stranger at a bar or to our aunt at a family party. I am just fortunate to have the chance not only to talk about science, but also to show people how it works on the streets where they live. ■

Carolyn Beans *is a biology graduate student at the University of Virginia in Charlottesville.*

BUZZ OFF

Contact has been made.

BY JOHN GRANT



Out of the bowels of space they came, the myriad ships of the greatest exploratory force the galaxy had ever seen, their sleek hulls glinting in the furtive starlight.

The Sgrin'th fleet eased gently into orbit around the blue, watery third planet of a moderate-to-small yellow star, reflector screens raised to make the arrival undetectable to the instrumentation of those on the world below. For the Sgrin'th had been able to tell from many light years away that this was indeed the home of a technological civilization — although not, of course, whether it would survive long enough still to be there when the fleet arrived. So many technological civilizations foundered young.

This one, however, was still extant — although only just. Cultural infantilization and deteriorating climate, the two deadly coupled factors that had accounted for so many civilizations, were well under way. The Sgrin'th preened themselves that they had arrived in time to save this one. They had been able to save many in similar situations before.

Those successes made the pain of the rare failures easier to bear.

The Sgrin'th knew better than to reveal immediately to the inhabitants of this world the glory of the interstellar fleet. Too many bellicose cultures would respond with pointless use of weaponry. Others would sink into apathy at the sight of technologies so very much in advance of their own.

There had also been the unique case of the nnHHptuths of Mondriodo XII, but the Sgrin'th never talked about *that* humiliating experience. Supernovae can happen for a diversity of later undiagnosable reasons.

To avoid future such unfortunate incidents, the method the Sgrin'th had devised over some billions of years was to send down individual emissaries to make telepathic contact with individuals in the highest echelons of power, so that it would be the aboriginals, not the Sgrin'th, who made the first steps towards clearing up the mess and graduating to membership of the Galactic Fellowship.

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

The Sgrin'th expedition leader, Nuit, called the appointed volunteer emissaries to the command bridge. Through the curved plexiglass viewports they could see the crowds of Sgrin'th ships hanging in hidden space.

"The aboriginals seem," said Nuit, "an ideal species for salvage. Most certainly they offer no threat to the rest of the Fellowship, and perhaps they could contribute much to the welfare of our galaxy-wide community.

"Even so, they might be dangerous."

What the fleet had been able to observe from orbit — able to observe with no great difficulty, in fact — was that the individuals of this planet's dominant species were far larger than Sgrin'ths. This was nothing new. The forces of evolution being what they are, some intelligent creatures are bigger or smaller than others. The nnHHptuths of Mondriodo XII, for example … but, wait, we don't want to talk about *them*.

"The atmosphere," continued Nuit, "is easily breathable by our kind, and none of the planet's microorganisms represent a threat. There seems no reason at all why our expedition to this world should be anything other than a resounding success!"

Nuit punched the air with a mandible.

The air in the command bridge was filled with cheers.

A dozen orbits later, everything was different.

The command bridge had about it the gloom of a mortuary.

Of the original 4,096 emissaries sent planetside, exactly four still lived.

"I don't understand it!" wailed Nuit. "We have never before met with such hostility!"

Nuit called up on the memory screen a typical encounter between a Sgrin'th emissary and one of the aboriginals — or humans, as they called themselves. The four emissaries clustered around Nuit knew already what they'd see, but their leader felt the need to show them again in the hope that someone might have a bright idea on how to deal with the crisis.

On the screen, they saw a young Sgrin'th approach a far larger human.

"We come to help you," telepathed the Sgrin'th by way of introduction.

The human picked up from its desk an artefact that was, in the local lingo, called a book.

There was a horrific squelching noise, and the picture faded to black.

"Is there any rational explanation?" cried Nuit into the silent echo of that terrible sound. "Never in all our billions of years of galactic exploration have we encountered such immediate hostility as this! Can anyone give me good cause why we shouldn't …? Well, I'm not mentioning Mondriodo here, but you can surely understand my drift."

No one had a reply.

"It's just so," said Nuit, wriggling its long, pointed, yellow-and-black-striped abdomen and buzzing its wings vexedly, "so *unreasonable*!" ■

John Grant *is the author of more than 60 books, both fiction and nonfiction — the latter including such works as* Discarded Science, Denying Science *and (with John Clute)* The Encyclopedia of Fantasy. *He has won the Hugo (twice), the World Fantasy Award and a bunch of other awards.*